AD________________

GRANT NUMBER DAMD17-96-1-6172

TITLE: Mechanism of Splicing of Unusual Intron in Human Proliferating Cell Nucleolor P120

PRINCIPAL INVESTIGATOR: Dr. Hongxiang Liu

CONTRACTING ORGANIZATION: Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724

REPORT DATE: December 1997

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 1997	**3. REPORT TYPE AND DATES COVERED** Annual (1 Dec 96 - 30 Nov 97)

4. TITLE AND SUBTITLE Mechanism of Splicing of Unusual Intron in Human Proliferating Cell Nucleolar P120	**5. FUNDING NUMBERS** DAMD17-96-1-6172
6. AUTHOR(S) Dr. Hongxiang Liu	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Cold Spring Harbor Laboratory Cold Spring Harbor, NY 11724	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Commander U.S. Army Medical Research and Materiel Command Fort Detrick, Frederick, Maryland 21702-5012	**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited	**12b. DISTRIBUTION CODE**

13. ABSTRACT *(Maximum 200*

In the past year I have concentrated on the mechanisms of recognition of splicing enhancers, which are relevant to both conventional and non-conventional intron splicing. Three novel classes of exonic splicing enhancers (ESEs) recognized by human SF2/ASF, SRp40 and SRp55 have been identified by an iterative functional selection procedure. These ESEs are functional in splicing and are highly specific. In most cases,only the cognate SR protein can efficiently recognize an ESE and activate splicing. An interesting exception is that SRp40-selected ESEs can function with either SRp40 or SRp55. UV-crosslinking/competition and immunoprecipitation experiments showed that SR proteins recognize their cognate ESEs in nuclear extract by direct and specific binding. A motif search algorithm was used to derive consensus sequences for ESEs recognized by each SR protein, and to show that such consensus sequences occur at high frequencies in exonic regions, particularly those corresponding to naturally occurring, mapped ESEs. Multiple high score motifs were also found in the proliferating cell nucleolar antigen (P120) gene exons, including those adjacent to the non-conventional intron F. Future studies will focus on the identification of splicing factors essential for the function of splicing enhancers, either in the context of conventional or non-conventional introns.

14. SUBJECT TERMS Breast Cancer P120, Splicing, U11/U12, Biochemistry, Gene Expression	**15. NUMBER OF PAGES** 41
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

HXL In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

HXL In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

HXL In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Han Xiang Liu 12/30/97

PI - Signature Date

TABLE OF CONTENTS

INTRODUCTION

Intron F of the P120 gene belongs to the minor class of non-conventional introns, which have non-canonical splice site and branch site sequences. Co-transfection experiments in cultured CHO cells indicated that splicing of this non-conventional intron required U12 snRNA, which is not required for conventional intron splicing (Hall and Padgett, 1995 Cold Spring Harbor Laboratory RNA Processing Meeting, published in 1996). I previously proposed to study the biochemical mechanism of this novel class of intron splicing (August 1995). At the time the fellowship was granted (December 1996), however, significant progress had already been achieved in this field. In vitro studies in Hela nuclear extract revealed that compared to conventional intron splicing, non-conventional intron splicing required common (U5) and unique (U11, U12, U4atac and U6atac) snRNPs (Tarn and Steitz, 1996a, 1996b; Yuo and Steitz, 1997). U11, U12, U4atac and U6atac functioned in the non-conventional spliceosome as analogues of U1, U2, U4 and U6, respectively, in the conventional spliceosome (Hall and Padgett, 1996; Tarn and Steitz, 1996a, 1996b; Kolossova and Padgett, 1997; Yuo and Steitz, 1997). SnRNPs U1, U2, U4 and U6 were shown not to be required for the splicing of non-conventional introns and are in fact not present in the non-conventional spliceosome. In some cases, depletion of U1 or U2 snRNA from the extract was necessary for detecting the splicing of a non-conventional intron (Tarn and Steitz, 1996). An interesting discovery was made in my mentor's lab, i.e., that there is cross-talk between adjacent conventional and non-conventional introns of the same gene (Wu and Krainer, 1996). In the in vitro splicing system, splicing of the non-conventional intron 2 of the sodium channel gene (which has the same elements present in the P120 intron F but splices more efficiently in vitro) was greatly stimulated if the downstream exon 3 was followed by the 5' splice site from the conventional intron 3. This stimulation was dependent on U1 snRNP, an snRNP only required for conventional intron splicing (Wu and Krainer, 1996). The cross-talk between these two distinct intron splicing pathways is likely a reflection of exon definition (Robberson et al., 1990) and is most probably mediated by exon splicing enhancers (ESEs), as previously reported for adjacent conventional introns (see below). Recent unpublished experiments from our lab have indeed shown that purine-rich exonic splicing enhancers can activate splicing of non-conventional introns. Therefore, I concentrated my effort in the past year to understand the molecular basis of exonic splicing enhancer function and their specific recognition by trans-acting factors, the SR proteins.

Three major aspects of the exon sequence have been shown to be relevant to splice-site selection: (i) the strength of the splice sites (Brunak and Engelbrecht, 1991); (ii) the length of the exon (Dominski and Kole, 1991; Sterner and Berget, 1993); and (iii) positive and negative exon *cis*-elements (Lavigueur et al., 1993; Sun et al., 1993; Tian and Maniatis, 1993; Watakabe et al., 1993; Xu et al., 1993; Caputi et al., 1994; Dirksen et al., 1994; Tanaka et al., 1994; Tian and Maniatis, 1994; Tsukahara et al., 1994; Amendt et al., 1995; Humphrey et al., 1995; Ramchatesingh et al., 1995; Staffa and Cochrane, 1995; Gontarek and Derse, 1996; Zheng et al., 1996). The positive exon *cis*-elements, known as exon splicing enhancers (ESEs), are often, though not always, found in a purine-rich context. A well-studied example is the ESE in the alternative exon M2 of the mouse IgM gene. This 73 nt ESE is essential for splicing of the preceding intron between exons M1 and M2. The M2 ESE can also stimulate splicing of the heterologous regulated intron of the *Drosophila doublesex* gene. Enhancer activity in the context of the IgM pre-mRNA could also be obtained by inserting certain natural or synthetic purine-rich sequences in place of the natural ESE. However, deletion of the purine-rich sequences within the M2 ESE did not abolish its activity completely (Watakabe et al., 1993; Tanaka et al., 1994). In agreement with this finding, SELEX experiments revealed that certain non-purine-rich sequences can also function as enhancers (Tian and Kole, 1995; Coulter et al., 1997). Most natural ESEs have been identified in tissue-specific or developmentally regulated exons, which typically have weak splice sites and require the ESE for exon inclusion. In some cases ESEs are specifically recognized by one or more SR proteins (Lavigueur et al., 1993; Sun et al., 1993; Tian and Maniatis, 1993; Tian and Maniatis, 1994; Ramchatesingh et al., 1995; Gontarek and Derse, 1996). In turn, SR proteins are expressed at different levels in different tissues, and their expression also appears to be regulated by alternative splicing (Jumaa et al., 1997; for review, see Cáceres and Krainer, 1997).

The SR proteins are a family of highly conserved serine/arginine-rich RNA-binding proteins. They are essential splicing factors (Krainer et al., 1990b; Ge et al., 1991; Krainer et al., 1991; Zahler et al., 1992) and also regulate the selection of alternative splice sites in a concentration-dependent manner (Ge and Manley, 1990; Krainer et al., 1990a; Zahler et al., 1993a), in part by antagonizing the activity of hnRNP A1 (Mayeda and Krainer, 1992). The SR proteins act very early in spliceosome assembly (Krainer et al., 1990a; Fu, 1993; Staknis and Reed, 1994). They promote the binding of U1 snRNP to the 5' splice site (Eperon et al., 1993; Wu and Maniatis, 1993; Kohtz et al., 1994; Staknis and Reed, 1994; Zahler and Roth, 1995) and of U2AF^{65} to the 3' splice site (Wu and Maniatis, 1993), apparently by interacting with U1 70K and U2AF^{35}, respectively. These observations have led to the hypothesis that SR proteins bound to ESEs recruit splicing factors to bind to the splice sites of adjacent introns (Wu and Maniatis, 1993; Staknis and Reed, 1994).

Nine human SR proteins are presently known: SF2/ASF, SC35, SRp20, SRp40, SRp75, SRp55, 9G8, SRp30c, and the somewhat more divergent p54. These proteins are closely related in primary structure and share the ability to complement splicing in a HeLa cell S100 extract (Ge et al., 1991; Krainer et al., 1991; Fu et al., 1992; Zahler et al., 1992; Zahler et al., 1993b; Cavaloc et al., 1994; Screaton et al., 1995; Zhang and Wu, 1996). SR proteins appear to have partially redundant functions, such that several different members of the family can complement an S100 extract to splice the same pre-mRNA, and/or stimulate use of the same alternative 5' splice sites *in vitro* or *in vivo* (Fu et al., 1992; Zahler et al., 1992). However, substrate-specific differences in general splicing, enhancer-dependent splicing, or alternative splicing mediated by different SR proteins have also been reported (Fu, 1993; Sun et al., 1993; Zahler et al., 1993a; Cáceres et al., 1994; Wang and Manley, 1995). More importantly, *Drosophila* SRp55/B52 has been shown to be essential for development (Ring and Lis, 1994; Peng and Mount, 1995), whereas at least one copy of the chicken SF2/ASF gene is required for survival of a B-lymphocyte cell line (Wang et al., 1996). Individual SR proteins also differ in their subnuclear localization signals and in their ability to shuttle between the nucleus and the cytoplasm (Cáceres et al., 1997a; Cáceres et al., 1997b). Finally, individual SR proteins exhibit striking phylogenetic sequence conservation of all their constituent domains (Birney et al., 1993). Taken together, these observations demonstrate that individual SR proteins have some unique, specific functions.

Although SR proteins have been clearly implicated in ESE recognition and function, predictive rules for the recognition of ESEs by different SR proteins have not been derived. In this study, I sought to determine the specificity of individual SR proteins in ESE recognition by performing a randomization and selection procedure under splicing conditions.

BODY

1. Experimental procedures

Preparation of HeLa cell extracts and recombinant SR proteins

Nuclear and cytosolic S100 extracts were prepared from fresh 12 l suspension cultures of HeLa cells, as described (Mayeda and Krainer, 1997b).

Expression and purification of the authentic form of the recombinant SR proteins SF2/ASF, SRp40 and SRp55, using the expression vector pET9c (Novagen), were carried out as described previously (Krainer et al., 1991; Screaton et al., 1995). The integrity and purity of these recombinant SR proteins were checked by SDS-PAGE and their specific activities were determined by *in vitro* splicing of β-globin pre-mRNA in S100 extract (data not shown).

Randomization and selection

The SELEX procedure is outlined in Figure 1A. The sequence of the wild type IgM exon M2 is shown in Figure 1B. The plasmids μM1-2 and μMΔ, which bear a mouse IgM minigene with or without the natural enhancer, respectively (Watakabe et al., 1993), were a generous gift from Prof. Y. Shimura. The randomized substrate pool was constructed by overlap-extension PCR (Horton et al., 1989; Tian and Kole, 1995). Two sets of PCRs were performed using μMΔ as template. The first PCR was carried out with primers R and A. The second PCR used primers P and B. The products from the two reactions were then combined and further amplified using primers P and A. The resulting PCR product was then used for *in vitro* transcription with T7 RNA polymerase to generate a radiolabeled pre-mRNA substrate pool. The pool of spliced mRNAs generated by *in vitro* splicing was excised from a urea-polyacrylamide gel, eluted in 0.5 M ammonium acetate plus 0.1% SDS, reverse transcribed using Superscript II RT (GIBCO-BRL) and amplified by PCR using primers P and A. The amplified product was further amplified using primers S and A. The PCR product was purified on a 2% agarose gel and re-assembled into the pre-mRNA template by overlap-extension PCR for the next round of selection. The reverse transcription and PCR reactions were performed as suggested by the vendors (GIBCO-BRL and Strategene, respectively). All the PCRs were done using the high fidelity Pfu DNA polymerase. The primers were purchased from Operon and were used at a concentration of 100 μM. After three rounds of selection, the amplified spliced products were subcloned into the vector PCR-Blunt (Strategene) and sequenced using a Dye Terminator Cycle Sequencing kit and an automated ABI 377 sequencer (Perkin-Elmer). The second and third rounds of SELEX were performed in nuclear extract depleted of total SR proteins by Mg^{2+} precipitation (Blencowe et al., 1994). The sequences of the primers were as follows:

Primer P: 5'-ATTTAGGTGACACTATAGAATAC-3'
Primer A: 5'-GCAGGTCGACTCTAGAAAGAAG-3'
Primer S: 5'-GTGAAATGACTCTCAGCAT-3'
Primer B: 5'-ATGCTGAGAGTCATTTCAC-3'
Primer R, 5'-GTGAAATGACTCTCAGCAT$(N)_{20}$CTAGTAAACTTATTCTTACGTC-3'

Identification of consensus motifs among the selected sequences

Functional selected sequences for each SR protein were aligned using Gibbs sampler (Lawrence et al., 1993), with the assumption that there is a common sequence motif of length L present at least once in all of the sequences. Since Gibbs sampler is a stochastic algorithm, for each fixed L, at least ten different runs (with different random seeds) were carried out for times sufficient to achieve convergence. A conservative value for L was determined empirically by observing a sharp drop in information per parameter (Lawrence et al., 1993) as L was increased. To exclude the possibility that the predicted consensus motif arose by chance, the information per parameter was also compared to alignments of random sequences obtained by shuffling the nucleotides within each sequence. The final alignment was manually adjusted in a few cases, when better matches to the consensus could be obtained by including a few flanking nucleotides in the alignment.

Construction of a scoring matrix

First, a frequency matrix f_i (a) was calculated from the alignment (i is the position of nucleotide a). Given a background frequency for the set of sequences, p(a), the scoring matrix is defined by the following formula:

$$s_i(a) = \log_2 \frac{f_i(a) + \varepsilon\, p(a)}{p(a)(1+\varepsilon)}$$

where i = {1, 2, ..., L) , a = {A, C, G, U), and $\varepsilon = 0.5$ is the Bayesian prior parameter (Lawrence et al., 1993).

A motif score is equal to the sum of the scores at each position. Motifs may be ranked by their scores. The top three scores in each sequence using all three different scoring matrices (SF2/ASF, SRp40, and SRp55) were calculated and tabulated (data not shown).

The sequence-scores were consistent semi-quantitatively with the gel intensity data when the sequence-scores for a given SR protein were defined as follows:

(a) The maximum score for each selected sequence was calculated using the scoring matrix, and the threshold was defined as the minimum of these scores.

(b) The sequence-score was defined as the number of non-overlapping motifs that have a score greater than or equal to the threshold. This integer score correlated well with the corresponding gel intensity.

It should be noted that a motif scoring matrix may depend on the pre-mRNA substrate and on the experimental parameters, such as the concentration of SR protein.

In vitro splicing

In vitro splicing was performed as described previously (Mayeda and Krainer, 1997a). Briefly, 20 fmol of ^{32}P-labeled, $^{7\text{CH}_3}$GpppG-capped SP6 or T7 transcripts generated from PCR products were incubated in 25 µl splicing reactions. The reactions contained 4 µl of HeLa nuclear extract or 7 µl of S100 extract in buffer D. The $MgCl_2$ concentration was 4.8 mM. 20 pmol of the appropriate SR protein was used in S100 complementation assays. After incubation at 30°C for 4 hours, the RNA was extracted and analyzed on 5.5% polyacrylamide denaturing gels, followed by autoradiography.

UV-crosslinking

UV-crosslinking experiments were carried out under splicing conditions with or without a 5-50 fold molar excess of unlabeled RNA competitor. Polyvinyl alcohol was omitted from the crosslinking reactions. After a 30 min incubation at 30 °C the reactions were exposed to 254 nm UV light using a Spectronics XL-1000 UV crosslinker at a setting of 1.8 J/cm^2 on ice. 10 µg of RNase A and 100 units of RNase T1 was added and the reactions were incubated for 15 min at 37 °C. The crosslinked proteins were analyzed by SDS-PAGE on a 12% gel, followed by autoradiography.

Immunoprecipitations

Immunoprecipitations were performed as described (Sun et al., 1993). The anti-SF2/ASF monoclonal antibody recognizes the N-terminus of SF2/ASF and does not crossreact with other human SR proteins (A.R. Krainer, unpublished). Polyclonal antiserum against SRp40 (anti-HRS/SRp40) was a generous gift from Drs. K. Du and R. Taub (Du et al., 1997). The crosslinking reactions were pre-cleared after incubation with control antibodies and 50 µl of protein A-agarose (1:1 suspension) in 500 µl of IP buffer (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 0.05% NP-40) for 2 hours at 4 °C. An unrelated monoclonal antibody of the same isotype was used for the SF2/ASF pre-clearing step and rabbit pre-immune serum was used for the SRp40 pre-clearing step. After spinning in a microcentrifuge for 30 min at 4 °C, the supernatants were transferred to tubes containing the appropriate antibody immobilized on protein A-agarose and rocked overnight at 4 °C. The bound material was recovered by centrifugation, washed twice with 1 ml of IP buffer, eluted in 30 µl of sample buffer (62.5 mM Tris-HCl, pH 6.8, 2% (w/v) SDS, 10% (v/v) glycerol, 5% (v/v) 2-mercaptoethanol), and analyzed by SDS-PAGE and autoradiography.

2. Results

Identification of SR protein target sequences from a random pool under splicing conditions

To find specific target sequences recognized by individual SR proteins under splicing conditions, a procedure based on SELEX (Tuerk and Gold, 1990) was utilized imposing a selection for splicing (Tian and Kole, 1995; Coulter et al., 1997), rather than for binding (Heinrichs and Baker, 1995; Tacke and Manley, 1995; Shi et al., 1997; Tacke et al., 1997). I further modified the procedure by carrying out the splicing reactions in the presence of a single, recombinant SR protein, which was used to complement HeLa extracts deficient for SR proteins – either an S100 extract or an SR protein-depleted nuclear extract. I chose to perform the selection for ESEs in the context of a well characterized IgM minigene transcript, comprising the last intron flanked by the M1 and M2 membrane isoform-specific exons. A prototypical ESE was previously mapped to a 73 nt fragment of exon M2 (Watakabe et al., 1993). This ESE was found to be essential for IgM pre-mRNA splicing in nuclear or S100 extract (Watakabe et al., 1993; our unpublished data). The scheme for the randomization and selection procedure is outlined in Figure 1A (see Materials and Methods for details). First, the natural ESE in the M2 exon was replaced by 20 nt of random sequence. A library of pre-mRNAs representing 1.2 × 10^{10} different sequences was spliced in S100 extract complemented by either recombinant SF2/ASF, SRp40, or SRp55. The spliced mRNAs, now carrying functional ESEs, were recovered from denaturing polyacrylamide gels. The randomized region of exon M2 of the spliced mRNAs was then amplified by RT-PCR and re-assembled into a new pool of pre-mRNAs for further selection. Two additional rounds of selection were carried out in SR protein-depleted nuclear extract (Blencowe et al., 1994) complemented with individual SR proteins, in order to mimic the conditions of nuclear extract, to minimize possible biases specific to the S100 extract, and to select the most efficient ESEs.

After three rounds of selection, the spliced mRNAs were amplified by RT-PCR, subcloned and sequenced. Twenty-four or more independent sequences obtained with each SR protein were analyzed to determine a consensus sequence, using the program GIBBS sampler (Lawrence et al., 1993). The defined motifs were used to generate a score matrix, according to the frequency of each nucleotide at each position. These score matrices were used to search the high-score motifs in each winner sequence. Small portions of the constant flanking regions (18 nt of the 5' region and 20 nt of the 3' region) were included during the search. The resulting alignments of sequences selected with SF2/ASF, SRp40, or SRp55 are shown in Figures 2A, 3 and 4, respectively. As a control, 30 sequences from the initial random RNA pool are shown in Figure 2B.

The consensus sequences derived for each of the three SR proteins tested differed in both length and sequence. Each of the consensus sequences is relatively degenerate, and not all of the individual selected sequences match the consensus at every position. However, many of the individual sequences have more than one good match to the consensus, allowing for one or two mismatches.

The SF2/ASF winners gave the consensus sequence SRSASGA (S represents G or C, R represents purine), which only in some cases corresponds to a purine-rich motif. The content of U residues in the SF2/ASF winner pool was 16%, which represents a significant reduction from the 21% of U residues found in the initial random pool. This reduction can be accounted for by the absence of U residues from the consensus motif. The content of C residues increased by 4%. The frequency of A and G did not vary significantly upon selection (Figure 2, A and B). SF2/ASF was previously shown to recognize purine-rich sequences in SELEX procedures based on binding; the reported sequences, RGAAGAAC and AGGACRRAGC (Tacke and Manley, 1995), are significantly different from, and simpler than, the consensus motif I found. Similar experiments, performed independently in our lab, revealed a different purine-rich consensus sequence, GARGAGC (A. Hanamura, I. Watakabe, and A.R. Krainer, unpublished data). In the present study, only 13 out of 28 winners have uninterrupted purine-rich motifs longer than 5 nt, indicating that SF2/ASF can productively recognize a far broader range of sequences. Indeed, the overall purine composition of the SF2/ASF-selected pool did not change significantly from that of the initial random pool.

The consensus for the SRp40-selected sequences is ACDGS (D represents residues other than C; S represents G or C). This consensus is also very different from that previously determined as an optimal RNA-binding site for SRp40, TGGGAGCRGTYRGCTCGY (Tacke et al., 1997). The content of G residues in the SRp40 winner RNA pool decreased from 39% in the initial random pool to 34%. The content of C residues increased by 5%. The frequency of A and U did not change significantly (Figure 3, 2B). The consensus motif for the SRp40 winners is relatively short but has a sufficient information content, such that, for example, it does not occur by chance in most of the RNAs sampled from the initial random pool. Winners SRp40-1, SRp40-2, SRp40-3, and SRp40-4, for example, are clearly not derived from a single founder sequence by accumulated mutations during PCR. However, they all share the sequence ACGGC, which matches the consensus, and is the only common sequence among these winners. Similar sequence relationships are seem among winners SRp40-5, SRp40-6, SRp40-7; SRp40-8, SRp40-9, SRp40-10; SRp40-11, SRp40-12; SRp40-13, SRp40-14; SRp40-15, SRp40-16.

The SRp55 winners yielded the consensus sequence USCGKM (S represents G or C; K represents U or G; M represents A or C). The C residue content in the SRp55 winner pool increased significantly, from 19% in the starting pool to 26%. The content of G and U residues decreased by 5% and 4%, respectively (Figure 4 and 2B). B52, which appears to be the *Drosophila* orthologue of human SRp55, was reported to have GRUCAACCNGGCGACNG as the optimal binding site (Shi et al., 1997). In that report, it was also suggested that a hairpin structure was required for efficient B52 binding. In contrast, I did not observe common secondary structure elements in our human SRp55 winner sequences.

The same sequence analysis programs were used to search the initial random pool, but no stable pattern was found. Each of the consensus sequences was used to create a score matrix. These score matrices were then used to search all three of the winner pools and the initial random pool. The mean score of the corresponding SR protein-selected winner pool was always higher than that of the other three pools (data not shown).

The SELEX winner sequences function as *bona fide* ESEs

To investigate the functional importance of the winner sequences, several winners were randomly chosen from the winner pools of each SR protein. Their ability to function as enhancers was tested by splicing the corresponding pre-mRNAs in HeLa nuclear extract or in S100 extract plus specific SR proteins (Figure 5).

All the SF2/ASF-selected sequences promoted efficient splicing in nuclear extract (Figure 5A, lanes 1, 4, 7, 10, 13, 16, 19, and 22), indicating that the selected sequences could function as true ESEs. Furthermore, these ESE sequences promoted splicing in S100 extract plus recombinant SF2/ASF (Figure 5A, lanes 3, 6, 9, 12, 15, 18, 21, and 24), but not in S100 extract only (Figure 5A, lanes 2, 5, 8, 11, 14, 17, 20, 23). The splicing efficiency in S100 extract plus SF2/ASF was lower than that of the nuclear extract. Winner sequences comprising either purine-rich (B1, B3, B4, B5 and B7) or non-purine-rich motifs (B2, B6, and B8) resulted in comparable splicing efficiencies.

A 20-40% ammonium sulfate cut of nuclear extract (NF20/40) was previously shown to be required for the function of synthetic ESEs (selected by a binding protocol) in S100 extract plus the appropriate SR protein (Tacke and Manley, 1995; Tacke et al., 1997). In contrast, the enhancers I selected by function were able to function in S100 extract plus the appropriate SR protein without further additions (Figure 5A). Supplementation with an NF20/40 fraction did not significantly improve the splicing efficiency (data not shown).

Similar results were obtained in splicing assays with the SRp40 winners. All of the 7 winners tested spliced in nuclear extract (Figure 5B, lanes 1, 4, 7, 10, 13, 16, and 19) and in S100 extract plus recombinant SRp40 (Figure 5B, lanes 3, 6, 9, 12, 15, 18, and 21), but not in S100 extract only (Figure 5B, lanes 2, 5, 8, 11, 14, 17, and 20). The splicing efficiency of the SRp40 winners in nuclear extract was lower on average than that of the SF2/ASF winners (Figure 5A, 5B). In S100 extract alone, several of the pre-mRNAs were extensively degraded (Figure 5B, constructs E1, E2, E4 and E7). This is consistent with the notion that SR proteins are involved in assembly of a commitment complex, such that substrates that are not productively assembled into commitment complexes and pre-spliceosomes are generally more susceptible to degradation by non-specific nucleases in the extract.

The eight tested SRp55 winners all spliced in nuclear extract (Figure 5C, lanes 1, 5, 9, 13, 17, 21, 25, and 29) and in S100 extract plus SRp55 (Figure 5C, lanes 4, 8, 12, 16, 20, 24, 28, and 32), but not in S100 extract only (Figure 5C, lanes 2, 6, 10, 14, 18, 22, 26, and 30). Interestingly, four of the eight SRp55 winners tested spliced more efficiently in S100 extract plus SRp55 than in nuclear extract alone (Figure 5C, C1, C4, C6 and C7), suggesting that SRp55 is the only SR protein required for effective recognition of these winner sequences. The higher splicing efficiency of C1, C4, C6 and C7 pre-mRNAs in S100 compared to nuclear extract cannot be accounted for by their increased stability, since the remaining winners, C2, C3, C5 and C7, were also greatly stabilized in S100 extract plus SRp55, but their splicing efficiencies were lower than in the nuclear extract.

I tested whether the short consensus motifs are sufficient to activate splicing. This was done by replacing sequences within template A13 by the short consensus motifs from the B1, B2, C4, or E7 winners (Figure 2A) and then testing the splicing activity of the corresponding pre-mRNAs. A13 was isolated from the initial random pool and had very low splicing activity in nuclear extract and no splicing activity in S100 extract complemented with any of the three SR proteins tested. Insertion of one copy of the consensus motifs was sufficient to activate splicing of the modified A13 pre-mRNA in S100 extract plus the cognate SR protein, although the splicing efficiencies were very low (data not shown). The low efficiency suggests that the sequence context surrounding the conserved motifs is also important for splicing, consistent with the observation that several of the winner sequences contain more than one match to the consensus. I also made and tested a number of clustered point mutations in several of the ESEs selected by each SR protein. I was unable to inactivate ESE function either by mutations in the best match to the consensus motif or by mutations on either side of the motif (data not shown). This unexpected observation suggests that each of the selected sequences has a high level of internal functional redundancy, which is probably necessary to allow efficient splicing.

SR protein specificity of the selected ESEs

The fact that each SR protein selected ESEs that fit a different consensus sequence suggests that SR proteins recognize the synthetic ESEs in a sequence-specific manner. On the other hand, the observation that most of the winner sequences promoted more efficient splicing in nuclear extract than in the S100 complementation reactions suggests that SR proteins may function cooperatively. I examined these possibilities by testing the effect of different SR proteins on splicing of pre-mRNAs with each type of winner. These experiments were performed in S100 extract complemented with individual SR proteins or with pairwise combinations thereof. Strikingly, the three kinds of SR protein winner sequences showed very different specificities.

The SRp40-selected winner sequences failed to splice in S100 extract plus SF2/ASF (Figure 6A, lanes 1, 4, 7, 10, 13, 16 and 19). However, they did splice in S100 extract plus SRp55 (Figure 6B, lanes 1, 4, 7, 10, 13, and 16). Adding two SR proteins together did not significantly increase the splicing efficiency, although additive effects were observed with two of the winners, E4 and E6, using SRp40 and SRp55 (Figure 6A, 6B).

The SF2/ASF-selected winner sequences gave a different result. The B2 winner spliced very poorly in the presence of S100 extract and SRp55, whereas the remaining winners, B1, B3, B4, B5, B6 and B7, failed to splice under these conditions (Table 1). SRp40 did not activate splicing of any of the SF2/ASF-selected winners tested. Moreover, SRp40 inhibited splicing of the SF2/ASF-selected winners even in the presence of SF2/ASF (Table 1).

The SRp55 winner C1 spliced in S100 extract plus any of the three SR proteins I examined. However, addition of two SR proteins did not increase its splicing efficiency. All the 6 other SRp55 winners tested failed to splice in S100 extract plus SF2/ASF or SRp40 (Table 1).

SR proteins bind specifically to the ESEs in nuclear extract

I have selected ESEs that respond specifically to individual SR proteins under splicing conditions. Because the selection was on the basis of function, it does not necessarily follow that the SR proteins bind directly to the cognate ESEs, although this is generally thought to be the case for at least some natural enhancers (see Discussion). To determine if SR proteins directly contact the novel ESEs, I carried out UV-crosslinking experiments under splicing conditions in nuclear extract. I used radiolabeled RNA fragments comprising the M2 exon with the different ESEs. 20 fmol of an M2 exon RNA comprising the SF2/ASF-selected winner B1 (referred to as B1E) was incubated in the presence of excess cold exon M2 RNA competitors with different ESEs. The reaction mixtures were then irradiated with UV light on ice, digested with RNases A and T1 and analyzed by SDS-PAGE. B1E crosslinked specifically to a 34-kDa polypeptide (Figure 7A). This crosslink could be competed by cold B1E RNA (lanes 2 and 3), but not by exon M2 RNAs with an SRp40- or an SRp55-selected ESE (lanes 10 and 11, and lanes 4 and 5, respectively). Exon M2 RNAs with ESEs selected by other SR proteins (lanes 8, 9, 12 and 13) or with sequences from the initial random pool (lanes 6, 7) also failed to compete with B1E for crosslinking to the 34-kDa polypeptide.

To confirm that the 34-kDa polypeptide is SF2/ASF, I carried out immunoprecipitations after UV crosslinking and RNase digestion (Figure 7B). As expected, the crosslinked 34-kDa polypeptide was immunoprecipitated by a monoclonal antibody specific for SF2/ASF (lane 2) but not by a control monoclonal antibody (lane 3).

Similar UV-crosslinking experiments were attempted using SRp40- and SRp55- selected ESEs. Crosslinked proteins of approximately 40 kDa and 55 kDa were detected using radiolabeled exon M2 RNAs corresponding to the SRp40 winner E7 (E7E), and the 55-kDa protein also crosslinked to the SRp55 winner C4 (C4E), although the background was high (data not shown). Neither of these RNAs crosslinked to proteins with the mobility of SF2/ASF. Crosslinking to the 55-kDa protein was competed by an excess of cold C4E RNA, but not by B1E or E7E. Immunoprecipitation with a polyclonal antiserum that recognizes both SRp40 and SRp55 (Du et al., 1997) selectively precipitated crosslinked proteins of the expected size (data not shown). These results suggest that, similar to SF2/ASF, SRp55 and SRp40 also interact directly with their cognate *in vitro*-selected ESEs.

Sequences that fit the consensus for *in vitro*-selected ESEs are present in natural exons and known ESEs

Sequences identified by SELEX procedures do not necessarily correspond to functional elements that have evolved in nature (Irvine et al., 1991). To evaluate the biological significance of the novel ESE consensus sequences I identified, I analyzed their distribution in known sequences of natural genes. I reasoned that if the short consensus motifs derived from the *in vitro*-selected ESEs are akin to natural ESEs, they should be present with higher probability in regions corresponding to known ESEs than elsewhere in the exons or in intron sequences. The score matrices derived for each of the three SR proteins tested were used to search genes or exons with previously characterized ESEs. The resulting scores were then plotted against the position along the exons or genes (Figure 8).

The natural sequence of the mouse IgM exon M2 was analyzed first, since our ESEs were selected in the context of this exon, after deletion of its natural ESE. Remarkably, a high density of motifs with high-score matches to the SF2/ASF and SRp40 consensus was found within the 73-nt natural ESE mapped previously (Watakabe et al., 1993). In contrast, few matching sequences were found in the flanking regions of the exon, and most of these had lower scores, correlating with the lack of splicing upon deletion of the natural ESE. The distribution of motifs with high-score matches to the SRp55 ESE consensus did not correlate with the location of the natural ESE. The SR protein specificity of the natural M2 ESE was not known from previous work, but I have determined that IgM minigene transcripts comprising the natural ESE can function in S100 extract complemented with SF2/ASF, SRp40 or SRp55 (data not shown). The high-score SF2/ASF and SRp40 motifs are present in clusters, suggesting that multiple copies of these motifs are particularly effective as ESEs, or provide an optimal context. Indeed, multimerization of short repeats often results in increased ESE activity, both in natural and synthetic enhancer elements (Tian and Maniatis, 1993; Tanaka et al., 1994; Tacke and Manley, 1995).

I next analyzed the sequence of the last exon of the bovine growth hormone (bGH) gene, which contains a natural ESE previously mapped to a 115-nt fragment, which is required for splicing of the preceding intron (Sun et al., 1993). The highest density of sequences matching the SF2/ASF, SRp40 and SRp55 consensus ESEs was found within the 115-nt fragment corresponding to the natural ESE, compared to the rest of the 189-nt exon. The highest scores for each of the three SR protein motifs were all found within the fragment with natural enhancer activity. Although the last intron of the bGH pre-mRNA does not splice in S100 extract in the presence of SR proteins, as it apparently requires additional factors, the ESE in the last exon was previously shown to bind SF2/ASF specifically, and this SR protein also stimulated bGH splicing in nuclear extract (Sun et al., 1993).

The caldesmon pre-mRNA is alternatively spliced in a tissue-specific manner (Humphrey et al., 1995). An alternative 5' splice site within the large exon 5 is used in non-muscle cells, which also exclude exon 6. In smooth muscle, the entire exon 5 is included and spliced to exon 6. A 32-nt repeat present in multiple copies within the 3' portion of exon 5 functions as an ESE to enhance usage of the upstream non-muscle-specific 5' splice site (Humphrey et al., 1995). Our sequence analysis showed that SF2/ASF and SRp40 ESE consensus sequences are highly enriched within the 3' portion of exon 5, whereas SRp55 consensus sequences are found much more frequently upstream of the non-muscle-specific 5' splice site.

Female-specific alternative splicing of the *Drosophila doublesex* (dsx) pre-mRNA involves six 13-nt repeat elements (dsxRE) and a purine-rich element (PRE) (Tian and Maniatis, 1993; Lynch and Maniatis, 1994). These *cis*-acting elements are essential for splicing of a dsx pre-mRNA in Hela cell nuclear extract. UV-crosslinking analysis showed that the dsxREs bind specifically to the human SR protein 9G8, whereas the PRE binds preferentially to SF2/ASF and probably other SR proteins of similar size in Hela nuclear extract (Lynch and Maniatis, 1996). Consistent with these results, our sequence analysis did not reveal any high-score motifs matching the SF2/ASF, SRp40, and SRp55 ESE consensus sequences within the dsxRE, whereas high-score matches to the SF2/ASF ESE were found within the PRE.

Finally, I also analyzed the sequences of characterized ESEs present in exon 5 of chicken cardiac troponin T (Xu et al., 1993), in exon 3 of the Tat gene of equine infectious anemia virus (Gontarek and Derse, 1996), in late pre-mRNAs of bovine papilloma virus type 1 (Zheng et al., 1996), in exon ED-A of human fibronectin (Lavigueur et al., 1993; Caputi et al., 1994), and in the exon downstream of the tat-rev intron of HIV-1 (Amendt et al., 1995; Staffa and Cochrane, 1995). In all cases, the sequence analysis was consistent with the available data on these natural ESEs and the binding of SR proteins, when known (data not shown).

Next I used the same score matrices to analyze the distribution of high-score motifs in human exons versus introns. 570 intron-containing genes, corresponding to 2634 exons (431 kb) and 2079 introns (1300 kb), were extracted from the ALLSEQ data (Burset and Guigo, 1996) and analyzed. I searched all sequences with a score equal to or greater than the mean score of the selected winner pool for each SR protein. Remarkably, high-score motifs matching each of the three SR protein ESE consensus sequences were found more frequently in exons than in introns. For SF2/ASF, the density of high-score motifs was 4.3/kb of exon and 2.9/kb of intron; for SRp40, the corresponding numbers were 7.9/kb of exon and 6.8/kb of intron; and for SRp55, they were 5.5/kb of exon and 4.9/kb of intron. The higher density of high-score motifs in exons than in introns is statistically significant because of the large database size, and the p-values for these pairwise comparisons were all less than 10^{-10}.

3. Discussion

I have developed a method to identify exonic splicing enhancer (ESE) elements that can function specifically with individual SR proteins. This was accomplished using SR protein-deficient HeLa extracts complemented with individual SR proteins, and a pool of pre-mRNAs derived from mouse IgM, whose natural ESE in exon M2 was replaced by a 20-nt segment of random sequence. I have successfully carried out this procedure with three human SR proteins – SF2/ASF, SRp40, and SRp55 – and have identified three novel classes of ESEs recognized and activated by these proteins. These ESEs are functional and specific: in most cases, each type of ESE activated splicing only in response to the SR protein in whose presence it was selected. The SRp40-selected ESEs responded both to SRp40 and to SRp55, indicating that some, though probably not all ESEs recognized by SRp40 represent a subset of ESEs recognized by SRp55. UV-crosslinking and immunoprecipitation experiments suggested that these SR proteins interact directly with their cognate ESEs through sequence-specific binding. Sequence analysis revealed that the motifs identified by the selection for function are present at much higher density in regions corresponding to known, natural ESEs than in other exon regions.

The initial randomized pool of IgM-derived substrates consisted of 20 fmol of pre-mRNA (~ 1.2×10^{10} molecules), which is large enough to include all possible 16-mers (~4.3×10^{9}). The longest motif I identified was the 7-nt consensus selected by SF2/ASF, indicating that the initial random pool had sufficient complexity. In parallel SELEX experiments I also employed a different RNA pool, in which only 14 positions within the IgM M2 exon were randomized. Functional ESEs were also selected out of that library (data not shown), suggesting that the 20-mer library can potentially encode most, if not all, natural ESEs, and that the library size I used was adequate. The functional SELEX procedure was performed for only three rounds. This was deemed sufficient, since all the winner sequences tested proved to be functional. Additional rounds of selection would be expected to result in loss of consensus sequence information, as only the most efficiently spliced RNAs would be recovered.

My experiments confirm and extend two previous studies that used functional selection from random pools to identify novel ESEs. Tian and Kole (1995) randomized a 20-nt region within the context of a duplicated exon in a model β-globin pre-mRNA. They selected sequences that promoted inclusion of the duplicated exon in HeLa nuclear extract. The resulting ESEs after five or seven selection cycles included both purine-rich and non-purine-rich motifs (Tian and Kole, 1995). A related approach was used by Coulter *et al.* (1997) to identify ESEs that promote inclusion of the alternative exon 5 of chicken cardiac troponin T. In this case, the natural ESE was replaced by a 13-nt randomized segment, and the selection for splicing was carried out by three rounds of transient transfection into QT35 quail cells. The resulting ESEs included both purine-rich elements and a novel class of AC-rich elements (ACEs)(Coulter et al., 1997). These pioneering studies could not readily identify the factors responsible for ESE recognition – although SR proteins were obvious candidates – because they relied on crude nuclear extracts or cultured cells. In addition, the novel ESEs did not fall into obvious consensus sequences, most likely because they represent a complex collection of elements recognized by several distinct factors. I improved this general approach by performing the selections in extracts dependent upon addition of individual SR proteins, which allowed us to identify functional ESEs recognized and activated by each SR protein, and therefore to derive a corresponding consensus sequence. Our selections were carried out in a different context from those in the above two studies, i.e., the last exon of the IgM pre-mRNA. Although ESEs can generally function in different contexts (Watakabe et al., 1993; Coulter et al., 1997), it is possible that the exon inclusion assays can also select relatively weak ESEs, as there is generally a fine balance between exon inclusion and exon skipping. On the other hand, the IgM selection, which relies on splicing versus no-splicing, may yield relatively potent ESEs.

Previous work also attempted to address the specificity of SR proteins in ESE recognition by using conventional SELEX procedures based on selection for high-affinity binding. The first such study was carried out with recombinant human His-tagged SF2/ASF or SC35 lacking the RS domain (to decrease non-specific binding due to the overall basic charge) and nine cycles of selection by immobilized metal affinity chromatography and amplification (Tacke and Manley, 1995). In the case of SF2/ASF, the selection yielded purine-rich motifs. Some of the selected sequences could bind intact SF2/ASF and could function as SF2/ASF-specific ESEs when multimerized and placed in the context of a chimeric α-globin/NCAM pre-mRNA. In the case of SC35, although some of the selected sequences bound the intact protein with high affinity, they were unable to function as ESEs in the context of the α-globin/NCAM pre-mRNA. Similar experiments carried out in our laboratory using purified human SF2/ASF or SC35, a 20-nt random region, and four cycles of filter binding and amplification yielded degenerate consensus sequences that were somewhat different from those reported by others (A. Hanamura, I. Watakabe, and A.R. Krainer, unpublished data). High-affinity RNA-binding sites for human SRp40 were recently identified using His-tagged protein pre-incubated in S100 extract to allow phosphorylation by endogenous kinases. This procedure yielded an SRp40 consensus high-affinity binding site, which could function when multimerized and placed in the context of α-globin/NCAM pre-mRNA (Tacke et al., 1997). This ESE was SRp40-specific, but also required an additional fraction from nuclear extract to complement an S100 extract. The *Drosophila* ortholog of SRp55, B52, was also used to perform *in vitro* SELEX via nine rounds of filter binding and amplification. Conserved sequence and secondary structure motifs were suggested to be required for high-affinity binding (Shi et al., 1997).

Strikingly, the earlier SELEX experiments based on binding gave very different results from our current results using some of the same SR proteins but with selection cycles based on function. They also differed from a recent functional study of SC35, which yielded the consensus ESE motif TSCNGYY (T. Shaal and T. Maniatis, personal communication). First, the consensus sequences obtained by these two approaches are very different. The SF2/ASF motifs defined by binding appear to be a subgroup of those defined by function, although they yield relatively low scores when analyzed with our SF2/ASF score matrix. It should be noted that a purine-rich composition is not sufficient for function. Rather, specific sequences are required, since only some oligo-purine segments tested can function as ESEs in the context of the IgM pre-mRNA (Tanaka et al., 1994). Similarly, multiple transition mutations in the natural purine-rich cTNT ESE abolished its function (Ramchatesingh et al., 1995). Second, many of the winner sequences obtained by binding protocols were not functional as ESEs. In contrast, among the winner sequences obtained by our functional selection protocol, many were tested and all of these were functional ESEs. Third, the complexity of the winner sequences identified by binding SELEX is much lower than that of the ESEs identified by functional SELEX. As a result, the consensus sequences obtained from the binding selection are less degenerate than those I obtained through functional selection. This may be due in part to the use of more selection/amplification cycles in some of the binding SELEX experiments. In the natural situation, exon sequences are obviously very diverse. Degenerate sequence specificity is probably essential for a limited number of SR proteins to be able to recognize a very large number of ESE-containing exons in different genes.

The different results obtained by binding selection and functional selection protocols shed light into the mechanisms of ESE function. The binding selection is based on the affinity of RNA-protein interactions, and the iterative protocol is designed to yield the binding sites with the highest affinity for the protein of interest. However, it appears that the best binding sites are not necessarily the best functional sites, and in some cases a high affinity may preclude function. In addition, optimal interactions between an SR protein and its cognate ESEs may require other splicing components, as opposed to just the purified protein. The binding protocol is carried out with the purified protein, whereas the functional selection protocol is carried out in the presence of all components required for splicing. There are also technical reasons why iterative binding protocols may not yield optimal functional sites. The binding affinity and/or the specificity of the binding may be significantly affected by the idiosyncrasies of the binding assay employed. For example, in the most common binding assays, the electric field and electrophoresis buffer, or interactions with the nitrocellulose filter, the agarose resin, or the gel may influence binding (Irvine et al., 1991). Indeed, several of the SF2/ASF and SRp40 winners identified by iterative binding failed to bind to the cognate SR protein when analyzed by a different binding assay (Tacke and Manley, 1995; Tacke et al., 1997). Another contributing factor to the discrepancy between the results obtained in binding and functional assays may be that in at least some applications of the former, truncated proteins lacking the RS domain were used (Tacke and Manley, 1995). Although the precise functions of the RS domains are not completely understood, they appear to be important for protein-protein and/or RNA-protein interactions (Wu and Maniatis, 1993; Tacke et al., 1997; Xiao and Manley, 1997). Thus, deletion of the RS domain may affect the binding specificity, as may an incorrect or incomplete phosphorylation state of the domain. In addition, the use of oligo-histidine or other tags may also affect the binding specificity. Finally, in the case of the different consensus sequences obtained previously for *Drosophila* B52 (Shi et al., 1997) and in the present study for human SRp55, the binding specificity may have diverged considerably between arthropods and vertebrates. For example, the enhancer complex formed on the PRE of the *Drosophila doublesex* pre-mRNA binds SRp55/B52 in *Drosophila* Kc cell extracts, but does not appear to bind human SRp55 in HeLa cell extracts (Lynch and Maniatis, 1996).

For a given constitutively spliced pre-mRNA substrate, such as β-globin pre-mRNA, many SR proteins can individually support splicing in an S100 complementation assay (Fu et al., 1992; Zahler et al., 1992; Screaton et al., 1995; Zhang and Wu, 1996). This may reflect a partial overlap in the functions of the different SR proteins, or there may be multiple, redundant binding sites for the different SR proteins on certain pre-mRNAs. The striking phylogenetic conservation of each member of the SR family argues against their functional redundancy. Indeed, many examples of substrate-specificity differences among SR proteins have been described, such as in alternative splicing or commitment assays (Fu, 1993; Zahler et al., 1993a; Screaton et al., 1995; Wang and Manley, 1995). The fact that chicken SF2/ASF and *Drosophila* SRp55/B52 are essential for cell and embryo viability, respectively (Ring and Lis, 1994; Peng and Mount, 1995; Wang et al., 1996) argues that at least some functions important for development or cell viability are uniquely carried out by single SR proteins *in vivo*.

The specific recognition of ESEs by SR proteins is well documented. A purine-rich sequence in the ED-A exon of the fibronectin gene strongly enhances inclusion of this exon. Gel shift assays showed that this purine-rich sequence interacts specifically with SR proteins (Lavigueur et al., 1993). The last exon of the bovine growth hormone pre-mRNA has a purine-rich ESE required for splicing of the preceding intron; SF2/ASF binds specifically to this element in HeLa cell nuclear extracts and appears to be required for its function (Sun et al., 1993). The dsx repeat element (dsxRE) and purine-rich element (PRE) of the *doublesex* pre-mRNA bind to different SR proteins in both Hela and *Drosophila* Kc cell extracts. The PRE preferentially binds SF2/ASF in Hela cell extracts, and one of the SRp30 proteins in Kc cell extracts (Lynch and Maniatis, 1994; Lynch and Maniatis, 1996). The dsxRE binds 9G8 in Hela cell extracts and RBP1/SRp20 in Kc cell extracts (Lynch and Maniatis, 1994; Heinrichs and Baker, 1995; Lynch and Maniatis, 1996). The ESE of the alternatively spliced exon 5 of avian cardiac troponin T (cTNT) pre-mRNA binds to SF2/ASF, SRp40, SRp55 and SRp75 in Hela nuclear extract. Purified SRp40 and SRp55 can activate splicing of exon 5, but SC35 cannot (Ramchatesingh et al., 1995). The HIV tat exon 3 has a purine-rich ESE (Amendt et al., 1995; Staffa and Cochrane, 1995) and SF2/ASF but not SC35 can commit a tat minigene transcript to the splicing pathway (Fu, 1993; Chandler et al., 1997). Assembly of an enhancer complex (Enh complex, which resembles the commitment or E complex) *in vitro* results in recruitment of different SR proteins depending upon the ESE sequence, as judged by UV-crosslinking (Staknis and Reed, 1994). For example, the purine-rich bGH ESE crosslinked efficiently to SRp30 and at lower levels to SRp20 and SRp40 in the Enh complex. In contrast, the ESE of the avian sarcoma and leukosis virus (ASLV) crosslinked efficiently to SRp40 and poorly to SRp20 and SRp30, even though this ESE is also purine rich. Changing the purine-rich sequence of the ASLV ESE to a non-purine-rich sequence that retained ESE activity resulted in an increase in SRp30 crosslinking and a decrease in SRp40 crosslinking (Staknis and Reed, 1994).

My data address the molecular basis of the redundancy and specificity of SR proteins. The different consensus sequences of the three types of *in vitro*-selected ESEs, and their different responses to individual SR proteins provide an indication of the specificity of SR proteins in ESE recognition and function. The consensus sequence of SF2/ASF-selected ESEs, SRSASGA, matches the sequence of most purine-rich ESEs characterized to date. It is worthwhile to note that this sequence is devoid of U residues. In the HPRT and IgM genes, the presence of C residues within the purine-rich ESEs was compatible with enhancer function, whereas changing the C residues to U residues abolished their enhancer activity (Tanaka et al., 1994). The SRp55-selected winners also had a reduced U content, and I suspect that a low U composition contributes to the information content that defines ESEs recognized by these SR proteins.

The SRp40-selected ESEs share a relatively short consensus sequence, ACDGS. They could be activated by SRp55, but not by SF2/ASF. The fact that SRp55 could activate SRp40-selected ESEs suggests that these two SR proteins, which are closely related in domain structure, unmodified molecular mass (31.2 kDa for SRp40, 39.6 kDa for SRp55), and sequence (65% identity) (Screaton et al., 1995), also have some functional overlap. It is unlikely that the sequences selected by SRp40 fortuitously comprise a distinct SRp55 recognition site, but not an SF2/ASF site, since all of the seven independent SRp40 ESEs tested had similar properties. Using the SF2/ASF score matrix to search the seven examined SRp40-selected ESEs, most of them had a score lower than the minimum score of the SF2/ASF-selected ESEs, which could explain why SRp40-selected ESEs were not activated by SF2/ASF. I do not know why E4, which had a score higher than the average score of SF2/ASF-selected ESEs, was not activated by SF2/ASF. However, it is likely that the sequence context, e.g., in the form of negative elements or silencers, somehow prevents activation of this motif by SF2/ASF. A related observation was the fact that SRp40 inhibited splicing of some SF2/ASF-selected ESEs even in the presence of SF2/ASF. This may be due to formation of inhibitory complexes with SRp40, such that SR proteins may also participate in exonic silencer function, depending upon the sequence context. An interesting implication is that the variable expression levels of these antagonistic SR proteins may determine the cell type-specific function of certain ESEs.

I did not observe any cooperative effects among the SR proteins tested. However, the fact that most of the ESEs I identified gave higher splicing efficiencies in nuclear extract than in S100 extract complemented with SR proteins suggests that other SR proteins and/or additional splicing factors may be required for optimal ESE recognition or function. With other substrates, such as β-globin or certain natural ESE-dependent pre-mRNAs, comparable splicing efficiencies can be obtained in the two systems. Still other natural or synthetic ESE-dependent pre-mRNAs can only splice in the nuclear extract (Sun et al., 1993; Tacke and Manley, 1995). The ESEs I obtained were selected to function in S100 extract plus an SR protein, and hence, it is not surprising that at least basal function could be observed in this complementation system. However, maximal activity appears to require one or more additional factors that may be limiting in the S100 extract.

Many natural ESEs have been found in the last several years. Most of these well-defined ESEs are purine rich, although this nucleotide composition may reflect an experimental bias. First, many of the biochemical studies were carried out in Hela cell nuclear extract, in which SF2/ASF, which prefers purine-rich sequences, may be the most abundant SR protein. Second, purine-rich motifs may be easier to find by visual inspection which, together with the precedent of known purine-rich ESEs, makes them more likely to be studied further. I have identified three new degenerate motifs, which are not necessarily purine rich. Significantly, the SF2/ASF and SRp40 motifs I defined occur more frequently (and with higher scores) within exon segments corresponding to known ESEs than elsewhere in the exons. All of the motifs also occur more often in exons than in introns, and may thus contribute to defining exon-intron boundaries. These consensus sequences may be useful for the prediction of natural ESEs in uncharacterized exons. My data also suggest that target sites for multiple SR proteins are clustered within natural ESEs. This may explain why large deletions are often required to inactivate natural ESEs. The SRp55 ESE consensus motif did not always correlate with the location of natural ESEs. Interestingly, in the caldesmon gene, SF2/ASF and SRp40 sites are enriched in the 3' portion of exon 5 that is included in smooth muscle cells, whereas SRp55 sites occur more frequently in the 5' portion of exon 5 that is included in all cell types. Inclusion of the constitutively spliced upstream segment of exon 5 may also require a functional ESE, which I would predict, on the basis of the present data, to be SRp55-dependent. The differential recognition of the alternative 5' splice sites associated with the caldesmon exon 5 by different SR proteins may be responsible for the proper developmental and tissue-specific expression of caldesmon by alternative splicing.

CONCLUSION

In the past year I have concentrated on the mechanisms of recognition of splicing enhancers, which are relevant to both conventional and non-conventional intron splicing. Three novel classes of exonic splicing enhancers (ESEs) recognized by human SF2/ASF, SRp40 and SRp55 have been identified by an iterative functional selection procedure. These ESEs are functional in splicing and are highly specific. In most cases,only the cognate SR protein can efficiently recognize an ESE and activate splicing. An interesting exception is that SRp40-selected ESEs can function with either SRp40 or SRp55. UV-crosslinking/competition and immunoprecipitation experiments showed that SR proteins recognize their cognate ESEs in nuclear extract by direct and specific binding. A motif search algorithm was used to derive consensus sequences for ESEs recognized by each SR protein, and to show that such consensus sequences occur at high frequencies in exonic regions, particularly those corresponding to naturally occurring, mapped ESEs. Multiple high score motifs were also found in the proliferating cell nucleolar antigen (P120) gene exons, including those adjacent to the non-conventional intron F. Future studies will focus on the identification of splicing factors essential for the function of splicing enhancers, either in the context of conventional or non-conventional introns.

REFERENCES

Amendt, B.A., Z.-H. Si, and C.M. Stoltzfus. 1995. Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: Evidence for inhibition mediated by cellular factors. *Mol. Cell. Biol.* **15:** 4606-4615.

Birney, E., S. Kumar, and A.R. Krainer. 1993. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res.* **21:** 5803-5816.

Blencowe, B.J., J.A. Nickerson, R. Issner, S. Penman, and P.A. Sharp. 1994. Association of nuclear antigens with exon-containing splicing complexes. *J. Cell Biol.* **127:** 593-607.

Brunak, S., and J. Engelbrecht. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220:** 49-65.

Burset, M., and R. Guigo. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353-367.

Cãceres, J.F., and A.R. Krainer. 1997. Mammalian pre-mRNA splicing factors. In *Eukaryotic mRNA processing* (ed. A. R. Krainer), pp. 174-212. IRL press, Oxford.

Cãceres, J.F., T. Misteli, G.R. Screaton, D.L. Spector, and A.R. Krainer. 1997a. Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity. *J. Cell Biol.* **138:** 225-238.

Cãceres, J.F., G.R. Screaton, and A.R. Krainer. 1997b. A specific subset of SR proteins shuttles continuously between the nucleus and the cytoplasm. *Genes & Dev.*, in press.

Cãceres, J.F., S. Stamm, D.M. Helfman, and A.R. Krainer. 1994. Regulation of alternative splicing *in vivo* by overexpression of antagonistic splicing factors. *Science* **265:** 1706-1709.

Caputi, M., G. Casari, S. Guenzi, R. Tagliabue, R. Sidoli, C.A. Melo, and F.E. Baralle. 1994. A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon. *Nucleic. Acids Res.* **22:** 1018-1022.

Cavaloc, Y., M. Popielarz, J.P. Fuchs, R. Gattoni, and J. Stévenin. 1994. Characterization and cloning of the human splicing factor 9G8: a novel 35 kDa factor of the serine/arginine protein family. *EMBO J.* **13:** 2639-2649.

Chandler, S.D., A. Mayeda, J.M. Yeakley, A.R. Krainer, and X.-D. Fu. 1997. RNA splicing specificity determined by the coordinated action of RNA recognition motifs in SR proteins. *Proc. Natl. Acad. Sci. USA* **94:** 3596-3601.

Coulter, L., M. Landree, and T. Cooper. 1997. Identification of a new class of exonic splicing enhancers by *in vivo* selection. *Mol. Cell. Biol.* **17:** 2143-2150.

Dirksen, W.P., T.K. Hampson, Q. Sun, and F.M. Rottman. 1994. A purine-rich exon sequence enhances alternative splicing of bovine growth hormone pre-mRNA. *J. Biol. Chem.* **269:** 6431-6436.

Dominski, Z., and R. Kole. 1991. Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell. Biol.* **11:** 6075-6083.

Du, K., Y. Peng, L.E. Greenbaum, B.A. Haber, and R. Taub. 1997. HRS/SRp40-mediated inclusion of the fibronectin EIIIB exon, a possible cause of increased EIIIB expression in proliferating liver. *Mol. Cell. Biol.* **17:** 4096-4104.

Eperon, I.C., D.C. Ireland, R.A. Smith, A. Mayeda, and A.R. Krainer. 1993. Pathways for selection of 5' splice sites by U1 snRNPs and SF2/ASF. *EMBO J.* **12:** 3607-3617.

Fu, X.-D. 1993. Specific commitment of different pre-mRNAs to splicing by single SR proteins. *Nature* **365:** 82-85.

Fu, X.-D., and T. Maniatis. 1992. The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site. *Proc. Natl. Acad. Sci. USA* **89:** 1725-1729.

Fu, X.-D., A. Mayeda, T. Maniatis, and A.R. Krainer. 1992. General splicing factors SF2 and SC35 have equivalent activities *in vitro* and both affect alternative 5' and 3' splice site selection. *Proc. Natl. Acad. Sci. USA* **89:** 11224-11228.

Ge, H., and J.L. Manley. 1990. A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA *in vitro*. *Cell* **62:** 25-34.

Ge, H., P. Zuo, and J.L. Manley. 1991. Primary structure of the human splicing factor ASF reveals similarities with *Drosophila* regulators. *Cell* **66:** 373-382.

Gontarek, R.R., and D. Derse. 1996. Interactions among SR proteins, an exonic splicing enhancer, and a lentivirus Rev protein regulate alternative splicing. *Mol. Cell. Biol.* **16:** 2325-2331.

Hall, S.L., and Padgett, R.A. 1996. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns [see comments]. *Science* **271:** 1716-8.

Heinrichs, V., and B.S. Baker. 1995. The *Drosophila* SR protein RBP1 contributes to the regulation of *doublesex* alternative splicing by recognizing RBP1 RNA target sequences. *EMBO J.* **14:** 3987-4000.

Horton, R.M., H.D. Hunt, S.N. Ho, J.K. Pullen, and L.R. Pease. 1989. Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene* **77:** 61-68.

Humphrey, M.B., J. Bryan, T.A. Cooper, and S.M. Berget. 1995. A 32-nucleotide exon-splicing enhancer regulates usage of competing 5' splice sites in a differential internal exon. *Mol. Cell. Biol.* **15:** 3979-3988.

Irvine, D., C. Tuerk, and L. Gold. 1991. SELEXION. Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. *J. Mol. Biol.* **222:** 739-761.

Jumaa, H., J.L. Guenet, and P.J. Nielsen. 1997. Regulated expression and RNA processing of transcripts from the SRp20 splicing factor gene during the cell cycle. *Mol. Cell. Biol.* **17:** 3116-3124.

Kohtz, J.D., S.F. Jamison, C.L. Will, P. Zuo, R. Lührmann, M.A. Garcia-Blanco, and J.L. Manley. 1994. Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* **368:** 119-124.

Kolossova, I., and Padgett, R.A. 1997. U11 snRNA interacts in vivo with the 5' splice site of U12-dependent (AU-AC) pre-mRNA introns. *RNA* **3:** 227-33.

Krainer, A.R., G.C. Conway, and D. Kozak. 1990a. The essential pre-mRNA splicing factor SF2 influences 5' splice site selection by activating proximal sites. *Cell* **62:** 35-42.

Krainer, A.R., G.C. Conway, and D. Kozak. 1990b. Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells. *Genes & Dev.* **4:** 1158-1171.

Krainer, A.R., A. Mayeda, D. Kozak, and G. Binns. 1991. Functional expression of cloned human splicing factor SF2: homology to RNA-binding proteins, U1 70K, and *Drosophila* splicing regulators. *Cell* **66:** 383-394.

Lavigueur, A., H.L. Branche, R.K. Alberto, and B. Chabot. 1993. A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. *Genes & Dev.* **7:** 2405-2417.

Lawrence, C.E., S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootto. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262:** 208-214.

Lynch, K.W., and T. Maniatis. 1994. Synergistic interactions between two distinct elements of a regulated splicing enhancer. *Genes & Dev.* **9:** 284-293.

Lynch, K.W., and T. Maniatis. 1996. Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer. *Genes & Dev.* **10:** 2089-2101.

Mayeda, A., and A.R. Krainer. 1992. Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell* **68:** 365-375.

Mayeda, A., and A.R. Krainer. 1997. Preparation of Hela cell nuclear and cytosolic S100 extracts for *in vitro* splicing. *In Methods in Molecular Biology, RNA-Protein Interaction Protocols,* S.R. Haynes, ed., Humana Press Inc., Totowa, NJ. in press.

Min, H., C.W. Turch, J.M. Nikolic, and D.L. Black. 1997. A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer. *Genes & Dev.* **11:** 1023-1036.

Peng, X., and S.M. Mount. 1995. Genetic enhancement of RNA-processing defects by a dominant mutation in B52, the *Drosophila* gene for an SR protein splicing factor. *Mol. Cell. Biol.* **15:** 6273-6282.

Ramchatesingh, J., A.M. Zahler, K.M. Neugebauer, M.B. Roth, and T.A. Cooper. 1995. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol. Cell. Biol.* **15:** 4898-4907.

Ring, H.Z., and J.T. Lis. 1994. The SR protein B52/SRp55 is essential for *Drosophila* development. *Mol. Cell. Biol.* **14:** 7499-7506.

Robberson, B.L., Cote, G.J., and Berget, S.M. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* **10**: 84-94.

Screaton, G.R., J.F. Cáceres, A. Mayeda, M.V. Bell, M. Plebanski, D.G. Jackson, J.I. Bell, and A.R. Krainer. 1995. Identification and characterization of three members of the human SR family of pre-mRNA splicing factors. *EMBO J.* **14:** 4336-4349.

Shi, H., B.E. Hoffman, and J.T. Lis. 1997. A specific RNA hairpin loop structure binds the recognition motifs of the *Drosophila* SR protein B52. *Mol. Cell. Biol.* **17:** 2649-2657.

Staffa, A., and A. Cochrane. 1995. Identification of positive and negative splicing regulatory elements within the terminal tat-rev exon of human immunodeficiency virus type 1. *Mol. Cell. Biol.* **15:** 4597-4605.

Staknis, D., and R. Reed. 1994. SR proteins promote the first specific recognition of pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol. Cell. Biol.* **14:** 7670-7682.

Sterner, D.A., and S.M. Berget. 1993. *In vivo* recognition of a vertebrate mini-exon as an exon-intron-exon unit. *Mol. Cell. Biol.* **13:** 2677-2687.

Sun, Q., A. Mayeda, R.K. Hampson, A.R. Krainer, and F.M. Rottman. 1993. General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. *Genes & Dev.* **7:** 2598-2608.

Tacke, R., Y. Chen, and J.L. Manley. 1997. Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer. *Proc. Natl. Acad. Sci. USA* **94:** 1148-1153.

Tacke, R., and J.L. Manley. 1995. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J.* **14:** 3540-3551.

Tanaka, K., A. Watakabe, and Y. Shimura. 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol. Cell. Biol.* **14:** 1347-1354.

Tarn, W.Y., and Steitz, J.A. 1996a. Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns [see comments]. *Science* **273:** 1824-32.

Tarn, W.Y., and Steitz, J.A. 1996b. A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro. *Cell* **84:** 801-1.

Tian, H., and R. Kole. 1995. Selection of novel exon recognition elements from a pool of random sequences. *Mol. Cell. Biol.* **15:** 6291-6298.

Tian, M., and T. Maniatis. 1993. A splicing enhancer complex controls alternative splicing of *doublesex* pre-mRNA. *Cell* **74:** 105-114.

Tian, M., and T. Maniatis. 1994. A splicing enhancer exhibits both constitutive and regulated activities. *Genes & Dev.* **8:** 1703-1712.

Tsukahara, T., C. Casciato, and D.M. Helfman. 1994. Alternative splicing of β-tropomyosin pre-mRNA: Multiple cis-elements can contribute to the use of the 5'- and 3'-splice sites of the nonmuscle/smooth muscle exon 6. *Nucleic Acids Res.* **22:** 2318-2325.

Tuerk, C., and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249:** 505-510.

Wang, J., and J.L. Manley. 1995. Overexpression of the SR proteins ASF/SF2 and SC35 influences alternative splicing *in vivo* in diverse ways. *RNA* **1:** 335-346.

Wang, J., Y. Takagaki, and J.L. Manley. 1996. Targeted disruption of an essential vertebrate gene: ASF/SF2 is required for cell viability. *Genes & Dev.* **10:** 2588-2599.

Watakabe, A., K. Tanaka, and Y. Shimura. 1993. The role of exon sequences in splice site selection. *Genes & Dev.* **7:** 407-418.

Wu, J.Y., and T. Maniatis. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75:** 1061-1070.

Xiao, S.H., and J.L. Manley. 1997. Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes & Dev.* **11:** 334-344.

Xu, R., J. Teng, and T.A. Cooper. 1993. The cardiac troponin T alternative exon contains a novel purine-rich positive splicing element. *Mol. Cell. Biol.* **13:** 3660-3674.

Yu, Y.T., and Steitz, J.A. 1997. Site-specific crosslinking of mammalian U11 and u6atac to the 5' splice site of an AT-AC intron. *Proc Natl Acad Sci U S A* **12:** 6030-5.

Zahler, A.M., W.S. Lane, J.A. Stolk, and M.B. Roth. 1992. SR proteins: a conserved family of pre-mRNA splicing factors. *Genes & Dev.* **6:** 837-847.

Zahler, A.M., K.M. Neugebauer, W.S. Lane, and M.B. Roth. 1993a. Distinct functions of SR proteins in alternative pre-mRNA splicing. *Science* **260:** 219-222.

Zahler, A.M., K.M. Neugebauer, J.A. Stolk, and M.B. Roth. 1993b. Human SR proteins and isolation of a cDNA encoding SRp75. *Mol. Cell. Biol.* **13:** 4023-4028.

Zhang, W.J., and J.Y. Wu. 1996. Functional properties of p54, a novel SR protein active in constitutive and alternative splicing. *Mol. Cell. Biol.* **16:** 5400-5408.

Zheng, Z.-M., P. He, and C.C. Baker. 1996. Selection of the bovine papillomavirus type 1 nucleotide 3225 3' splice site is regulated through an exonic splicing enhancer and its juxtaposed exonic splicing suppressor. *J. Virol.* **70:** 4691-4699.

APPENDICES

Table legends

Table 1. Summary of the activities and specificities of three types of *in vitro*-selected ESEs.

The *in vitro*-selected ESEs were tested for function as part of IgM minigene pre-mRNAs in HeLa S100 extract complemented with recombinant SF2/ASF, SRp40, or SRp55. The sequences of the B, E, and C winner series are given in Figures 2, 3 and 4. ND: not determined.

Figure legends

Figure 1. Procedure for randomization and selection of exon splicing enhancers (ESEs).
(A) The natural ESE in mouse IgM exon M2 was replaced by a 20-nt segment of random sequence, and a library of pre-mRNAs was constructed by overlap-extension PCR and *in vitro* transcription. A sample of this pool, representing ~ 1.2 X 10^{10} pre-mRNA moleclules was then spliced *in vitro* by complementation of an S100 extract with individual recombinant SR proteins. The pool of spliced mRNA products was gel purified and the sequences corresponding to the ESE region were rebuilt into pre-mRNA template molecules for a new round of selection, or subcloned and sequenced. The sequences were analyzed by a motif-search algorithm to identify common patterns. (B) Sequence of the M2 exon of the mouse IgM gene. The sequence of the previously mapped ESE is shown in uppercase.

Figure 2. Sequence alignment of SF2/ASF-specific ESEs (A) and sequences from the initial random pool (B).
A consensus motif was identified as described in the text. The sequences are aligned on the basis of the best fit to the consensus within each sequence. Nucleotides in the boxed alignment that match the consensus position are shown white on black; mismatched nucleotides are not shaded. Underlined nucleotides are from the constant region flanking the randomized segment. The nucleotide composition of the randomized segment in the selected pool is provided in the lower left corner. S = G or C; R = A or G.

Figure 3. Sequence alignment of SRp40-specific ESEs.
D = A, G or U. See legend of Figure 2 for details.

Figure 4. Sequence alignment of SRp55-specific ESEs.
K = U or G; M = A or C. See legend of Figure 2 for details.

Figure 5. The selected sequences are functional SR protein-dependent ESEs.
(A) *In vitro* splicing of pre-mRNAs containing the SF2/ASF-selected winner sequences in HeLa nuclear extract (lanes 1, 4, 7, 10, 13, 16, 19, and 22), S100 extract alone (lanes 2, 5, 8, 11, 14, 17, 20, and 23) or S100 complemented by 20 pmol of recombinant SF2/ASF (lanes 3, 6, 9, 12, 15, 18, 21 and 24). (B) Splicing of SRp40-selected winner sequences in nuclear extract (lanes 1, 4, 7, 10, 13, 16 and 19), S100 extract alone (lanes 2, 5, 8, 11, 14, 17 and 20) or S100 extract complemented by 20 pmol of recombinant SRp40 (lanes 3, 6, 9, 12, 15, 18 and 21). (C) Splicing of SRp55-selected winner sequences in nuclear extract (lanes 1, 5, 9, 13, 17, 21, 25, and 29), S100 extract alone (lanes 2, 6, 10, 14, 18, 22, 26, and 30), or in S100 extract complemented by 10 pmol of SRp55 (lanes 3, 7, 11, 15, 19, 23, 27, and 31) or by 20 pmol of SRp55 (lanes 4, 8, 12, 16, 20, 24, 28 and 32). The structures and mobilities of the precursor, intermediates, and products of splicing (Watakabe et al., 1993) are shown next to each autoradiogram.

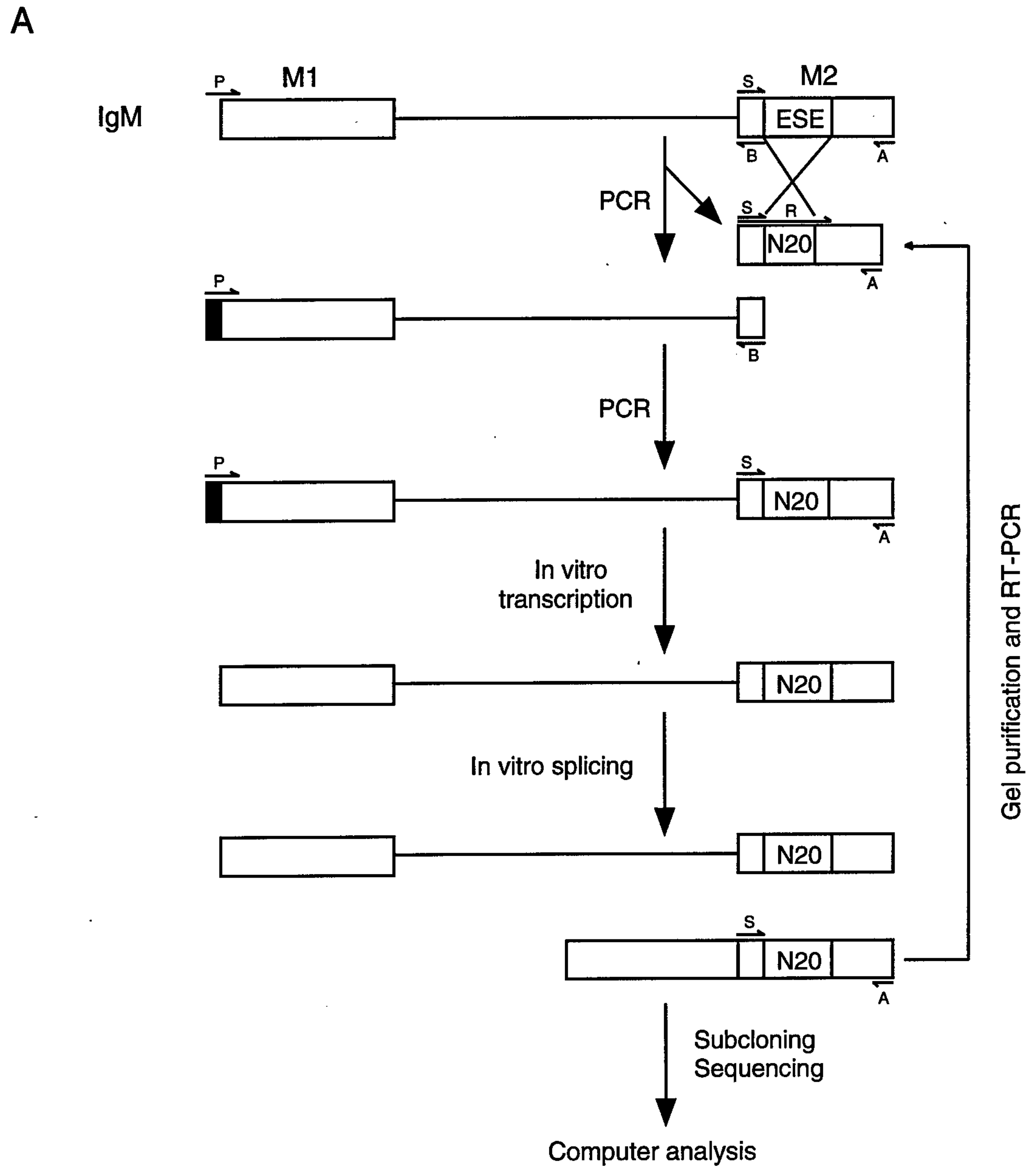
Figure 6. SR protein specificity of *in vitro*-selected ESEs.

(A) SRp40-selected ESEs are inactive with SF2/ASF. Splicing was carried out in HeLa S100 extract complemented with 20 pmol of SF2/ASF (lanes 1, 4, 7, 10, 13, 16 and 19), 20 pmol of SRp40 (lanes 2, 5, 8, 11, 14, 17 and 20), or 20 pmol of SF2/ASF plus 20 pmol of SRp40 (lanes 3, 6, 9, 12, 15, 18 and 21). (B) SRp40-selected ESEs can function in the presence of SRp55. Splicing was carried out in S100 extract complemented with 20 pmol of SRp55 (lanes 1, 4, 7, 10, 13, and 16), 20 pmol of SRp40 (lanes 2, 5, 8, 11, 14, and 17), or 20 pmol of SRp55 plus 20 pmol of SRp40 (lanes 3, 6, 9, 12, 15, and 18).

Figure 7. Specific binding of SF2/ASF to an SF2/ASF-selected ESE.
(A) UV-crosslinking competition binding assay. Radiolabeled exon M2 RNA (20 fmol) comprising the B1 winner sequence (B1E) was incubated under splicing conditions in HeLa nuclear extract. Subsequent UV-crosslinking and RNase digestion resulted in label transfer predominantly to two proteins of 34 kDa and 47 kDa. The former, which binds specifically, is presumed to be SF2/ASF (see below). Cold competitor RNAs containing either the B1 winner insert, an SRp55-selected insert (C4E), an SRp40-selected insert (E7E), or other control sequence inserts, were present in excess, as indicated above the autoradiogram. Lane 1: no competitor; in the remaining lanes, the indicated competitors were present in 5-fold excess (even lanes) or 50-fold excess (odd lanes) over the labeled B1E RNA. (B) Immunoprecipitation of SF2/ASF UV-crosslinked to the B1E RNA. UV-crosslinking was carried out as in panel A, lane 1. 5% equivalent of the input was loaded directly (lane 1). Parallel reactions were incubated with a control antibody (lane 3), or with anti-SF2/ASF monoclonal antibody (lane 2), and the immunoprecipitates were recovered in SDS gel loading buffer. In both panels, the samples were analyzed by SDS-PAGE and autoradiography.

Figure 8. Distribution of *in vitro*-selected ESE consensus sequences within exons comprising natural ESEs.
Score matrices were built for each class of *in vitro*-selected ESE, according to the frequency of each nucleotide at individual positions of each consensus sequence. The indicated natural exon sequences were searched with the score matrix, and the resulting scores (y axis) were plotted against the nucleotide positions for each exon (x axis). Note that the x axis scales are different in each case, because of the different exon sizes. Graphs are shown for mouse IgM exon M2, bovine growth hormone 3' exon, *Drosophila doublesex* female-specific exon, and chicken caldesmon exon 5. High score motif matches are shown by blue (SF2/ASF), red (SRp40), and yellow (SRp55) vertical bars. The green horizontal bars under the x axis indicate previously mapped ESEs or the *doublesex* purine-rich element (PRE). The black horizontal bars denote the *doublesex* repeat elements (dsxREs).

	SF2/ASF	SRp40	SRp55
B1	+++	+/-	+/-
B2	++	+/-	+
B3	+++	-	-
B4	+++	-	+/-
B5	+	+/-	+/-
B6	+++	+/-	+/-
B7	++	+/-	+/-
B8	+++	ND	ND
E1	-	+++	+++
E2	-	++	+++
E3	+/-	++	+
E4	-	++	++
E5	+/-	+	++
E6	-	+++	+++
E7	-	++	++
C1	++	++	+++
C2	-	-	+
C3	+/-	-	+++
C4	-	-	+++
C5	-	-	+
C6	+/-	-	++
C7	+/-	+/-	+++
C8	ND	ND	++



B

gtgaaatgactctcagcatGGAAGGACAGCAGAGACCAAGAGA
TCCTCCCACAGGGACACTACCTCTGGGCCTGGGATA
CCTGACTGTATGActagtaaacttattcttacgtctttcctgtgttgccctccag
cttttatctctgagatggtcttctttctagagtcgacctgc

SF2/ASF SELEX SEQUENCES

```
SF2/ASF-1                    CAAGCACAGUGACCGAGAAC
SF2/ASF-2                CGAUGUCCCGGAGGUUUUGC
SF2/ASF-3  (B1)           CGCGGUUAGGAGGAUGGAAA
SF2/ASF-4                        GGCACGGCGAGACACCAUCA
SF2/ASF-5                      GGCAGAGGAGAGCCGGGACGc
SF2/ASF-6                      GGCAGCGGGCGUACCCGGAU
SF2/ASF-7                        GGCACGGGGAGGCACCAUCA
SF2/ASF-8  (B2)               GGUCGCAGGUCAGGUGGGUU
SF2/ASF-9                        CCAGAGGGCGGAAACGUUGG
SF2/ASF-10          CGUGCCCACGUGUCUCAGGU
SF2/ASF-11          GGCUUGGUUCGCGGUGACGA
SF2/ASF-12             CGAUGACCCUCAGACGUAUA
SF2/ASF-13            GACGUCCAGUACGCUCGAGG
SF2/ASF-14                    CGCCGGACGACGUGUGUUG
SF2/ASF-15       UGAGUGCGCGGAUAGACUGACUA
SF2/ASF-16                    AUGCUCCGGAAUCGGAACGG
SF2/ASF-17 (B3)               GCGGACCCGGAAAGGACUAA
SF2/ASF-18                GUGGGUUCGGCGGAAUCAAG
SF2/ASF-19               GGAAGUACGGGACGUGCCGG
SF2/ASF-20 (B4)               CGUCGCAGGGCAGGUGGGAA
SF2/ASF-21                   AUCGGACAGGGUCCAGCAGG
SF2/ASF-22 (B5)      CGUGAAACUGCCCAGAGGUG
SF2/ASF-23 (B6)              AUAGGACUGGAUCGAGUUGG
SF2/ASF-24                     UUCGGACGGGCUAGGGAUGG
SF2/ASF-25 (B7)          GUUGCGGAGACGACCCGAGC
SF2/ASF-26                     AUCGGCCGAUCUGUGAGUUA
SF2/ASF-27 (B8)               CUCCAGACGUCGUUUGUUGC
SF2/ASF-28                    UGACAGCGGAAGGUACAGUG

CONSENSUS                        SRSASGA
```

G = 39% **A** = 22% **U** = 16% **C** = 23%

SEQUENCES OF INITIAL RANDOM POOL

R1		UCCUACGGUUGUUACCGGGA
R2		AGUGCGGUCACCGGAUGAGC
R3		UAUGACGAGCGGGAUCCGGG
R4		GUAGGCGUCUGGUGGGGGGG
R5		AUUCAGCCUAGUUGGGUGG
R6		CGUUAUACCGCGCCUGGGUG
R7		UCAGUGGAGGUUGUGGCACU
R8		GGGCCAUCGUUGUGGAGAAC
R9		UGGGCUCAGGCCGGCCGGUG
R10		CUCCUCGUUUAGGGGGUAGG
R11		GUGGGGUUCCGAUGGGGCCG
R12		AUAGCGGAUUACGGGCGGC
R13		GGGGAGGAGUUCGUGCUGAG
R14		GUCAUUAACGGACACAUGGC
R15		GUGAAUAUUGCGAUGUGAG
R16		CGUGAGUGAUUUCCACAACA
R17		CUUCAAGAUAGAACGUGGCU
R18	(A13)	AGACAGCGUGGGCGGGAGUG
R19		AGAGACAUCGAGGGACUAGG
R20		CACCGCGGUGCCACCUCCAC
R21		GAGAGACUGUUUUAGUACAC
R22		UGAGGACCAAAAGGGUGAAG
R23		UAGGGCGAGUAGUGAUAAUG
R24		UUGGCAUGCAGGAUAUGCGG
R25		AGUGCCUCGGUCAAACGGGG
R26		ACGAUCGGCAUGUCUUGUCG
R27		GGGGACGAAGCAAUAUGGGC
R28		UCGCAGACCAUCAAAUGCGG
R29		AGAUUUGCAGAUCGGUUGGA
R30		GAGGGAAGUAGAAAUGGCGC

G = 39% **A** = 21% **U** = 21% **C** = 19%

SRp40 SELEX SEQUENCES

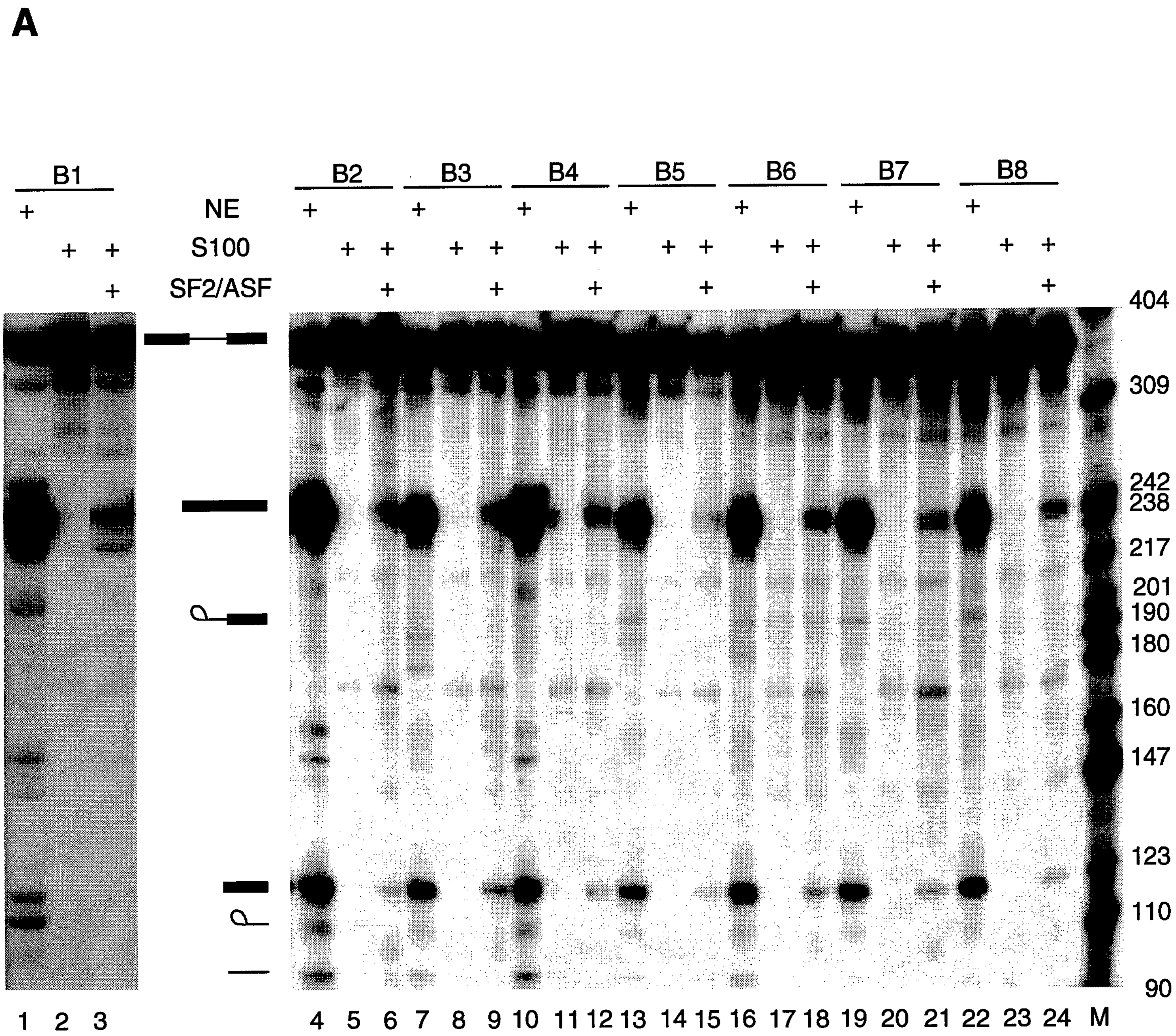
SRp40-1	UCUAAGGCGCUAAGA	ACGGC	
SRp40-2	AAC	ACGGC	UGUGAGUGGUCC
SRp40-3 (E4)	CGACGUGUGGGG	ACGGC	AAG
SRP40-4	CCAAUCGGAUCACCUA	ACGGC	
SRp40-5	GUAGG	ACUGG	AUCGCGUUGG
SRp40-6	CAGGGCACUUGUUUC	ACUGG	
SRp40-7	CCUC	ACUGG	ACUCAGUGGUG
SRp40-8	GUGAUACAU	ACAGG	UGGCGC
SRp40-9 (E5)	GGUAAGUACU	ACAGG	GUGUG
SRp40-10	GAAAGUUGUAAAG	ACAGG	GG
SRp40-11	ACAUGAACACAACG	UCGGG	G
SRp40-12	GGCGUUUUCGAGGA	UCGGG	A
SRp40-13	UGGAUGUCAGCG	ACGGG	CCA
SRp40-14 (E6)		ACGGG	CGGACUCCUCUGGUA
SRp40-15 (E3)	UUACA	ACUGC	ACCACGGUCG
SRp40-16	GCAGUG	ACUGC	AUUGGCAGC
SRp40-17	AGACCAGUAGCC	GCUGC	CGG
SRp40-18 (E1)	CGAGGAAUAU	AAAGG	UGGGA
SRp40-19 (E2)	AUGGGUCUG	ACACG	CUGACU
SRp40-20	ACGCUCAAUAGAAA	UCAAG	
SRp40-21	ACCAGGGUCGUCCG	UCUGG	G
SRp40-22	UGC	AGAGG	AUAGCCGGAACG
SRp40-23	CUUGAGGUGAAGG	UCAUG	UG
SRp40-24	GUAUUUCG	ACACC	AGUGUGA
SRp40-25	UGCUCACCCGGCCGCC	ACAGC	
SRp40-26 (E7)	UGU	GCAGC	UUGCGUCACGUC
SRp40-27	GAACCUU	GCAGG	UCGCGCGA
SRp40-28	UAGUA	ACCGC	GACAGUAGGC
SRp40-29	GACGUUGGUGUUA	UCCGC	CA
CONSENSUS		ACDGS	

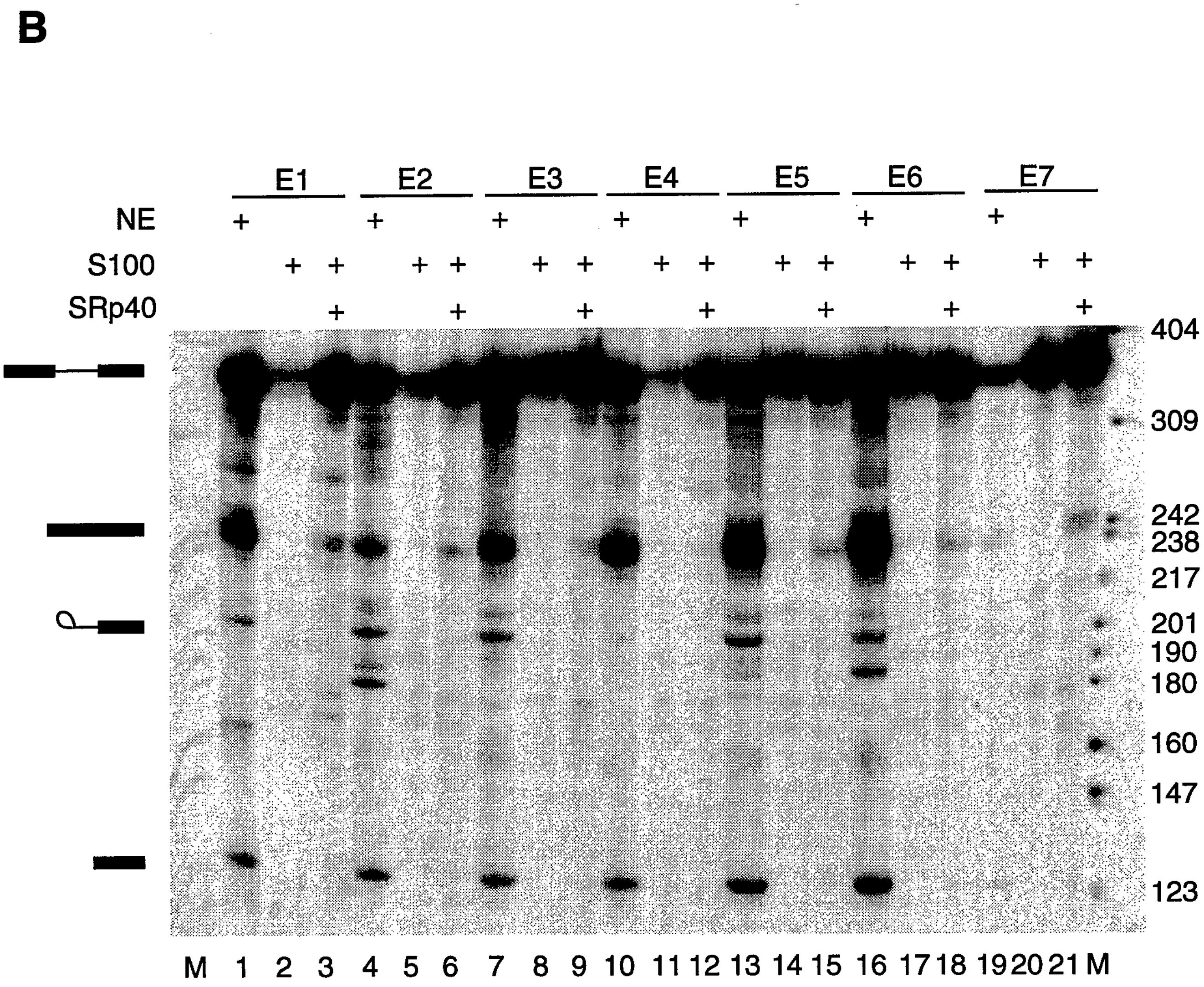
G = 34% **A** = 23% **U** = 19% **C** = 24%

SRP55 SELEX SEQUENCES

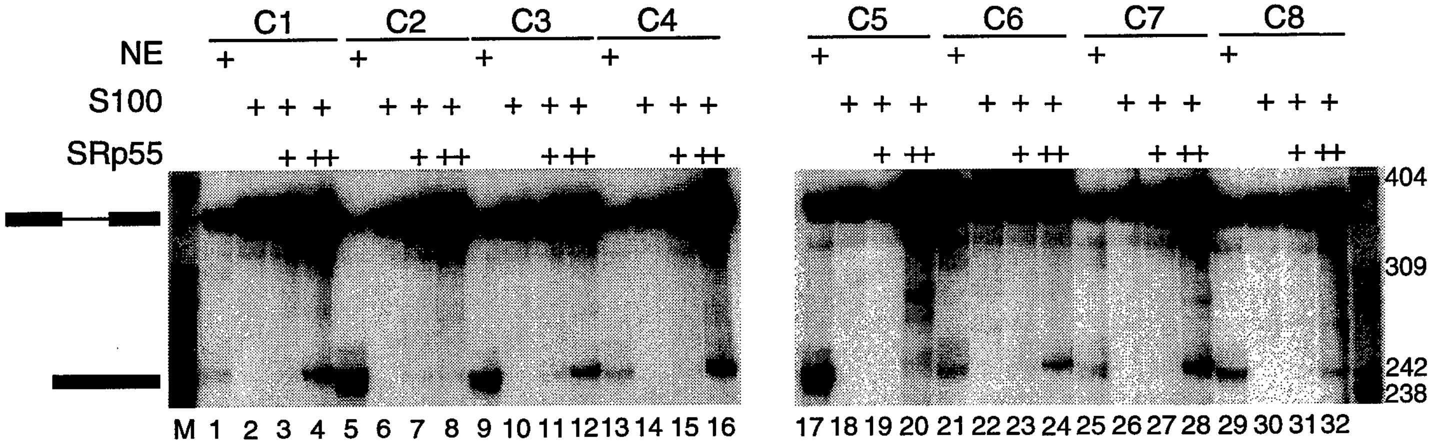
SRP55-1 (C1)		CGCGUC	GUGUCGUAGGGGGC
SRP55-2	AUGCAGACGAUGG	UGCGGC	U
SRP55-3	GAGUUGAGCGAUGG	UGCGUA	
SRP55-4 (C2)	AUAGCG	AGCGGA	AAACAGGUAA
SRP55-5 (C3)	CGAGC	CACGGA	CCACACGGA
SRP55-6	GGAUAACGGUGUGG	CCCGGC	
SRP55-7	G	GCCGGA	CGCAUUGCAGAG
SRP55-8 (C4)	UCCGAUCUGUG	CACGGA	CG
SRP55-9	AG	ACCGUC	AACAUGUCUGCC
SRP55-10 (C5)	CAAACC	UGCGUG	GUAUGGUA
SRP55-11	GG	ATCGUA	AGUGCAGACGA
SRP55-12	CAAACCGUCAAAG	UACGUC	A
SRP55-13	G	ACCGGG	AUUGAAGGAGCU
SRP55-14	CAUGAA	GCCGUC	ACCAACGUCUAG
SRP55-15	GGAUCGAA	UCCGGA	ACACGG
SRP55-16 (C6)	CGCACAC	UGCGUC	CCGGGGC
SRP55-17	GAUCGAA	UCCGGA	ACACAGG
SRP55-18		GACGUC	GCCCCGUGUGUAAG
SRP55-19 (C7)	CGUGU	CGCGUC	CUCGUGUGC
SRP55-20 (C8)	CACC	AGCGGA	GUCCCCAGAGC
SRP55-21	CGCG	UGCGUG	CAGUGCCAAG
SRP55-22	CGAUUCAGG	UACGUC	CAACU
SRP55-23	GG	AUCGUA	AGUGCAGAUGA
SRP55-24	GAUCGUA	UCCGGA	ACACGGG
CONSENSUS		USCGKM	

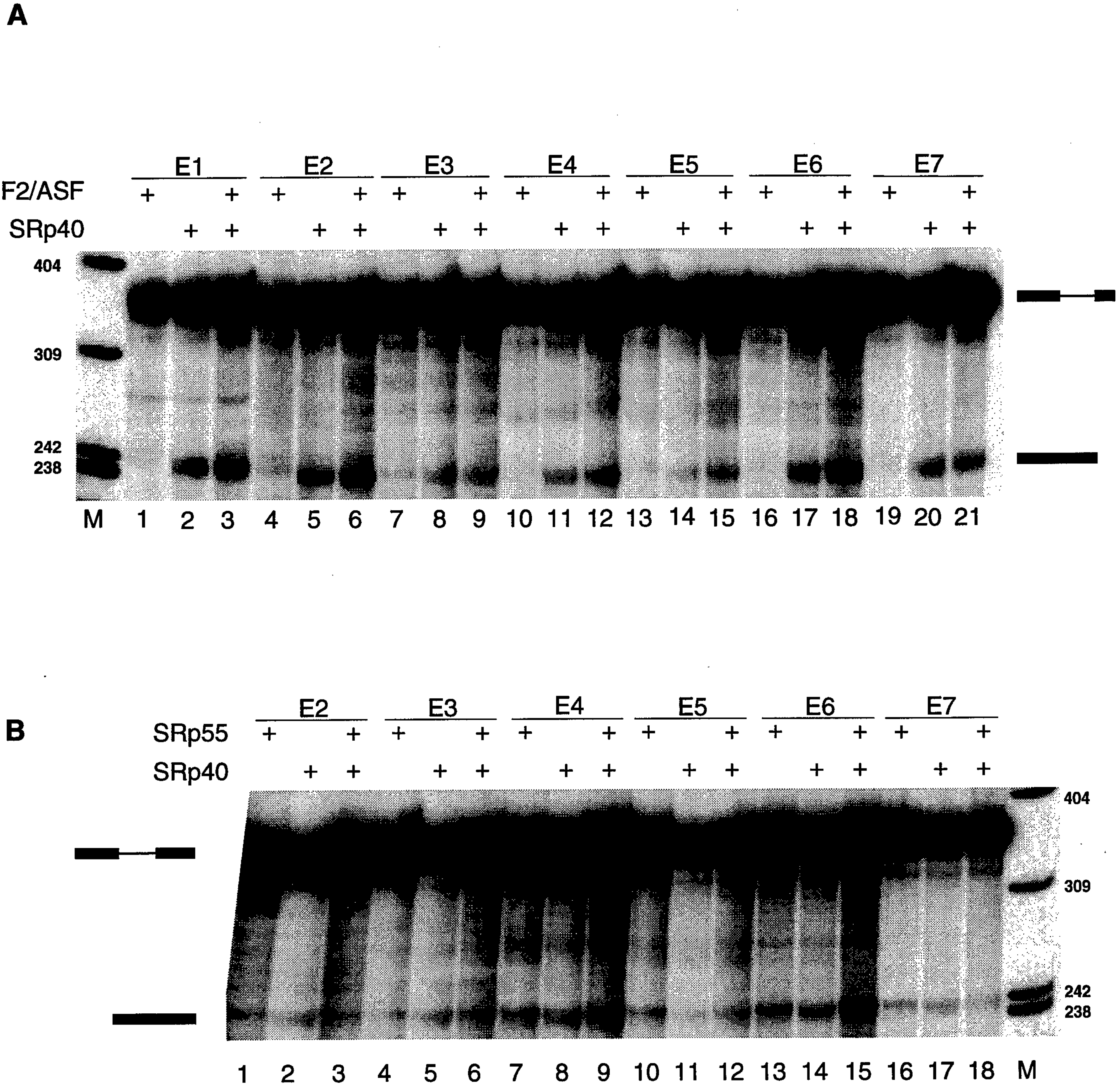
G = 34% **A** = 23% **U** = 17% **C** = 26%


A
B1
B2
B3
B4
B5
B6
B7
B8
NE
S100
SF2/ASF
404
309
242
238
217
201
190
180
160
147
123
110
90
1 2 3
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 M

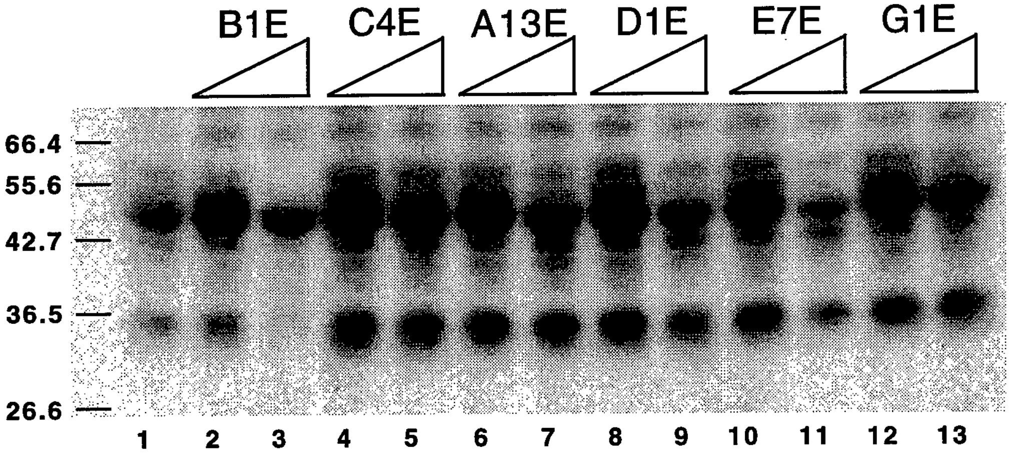

B
E1
E2
E3
E4
E5
E6
E7
NE
S100
SRp40
404
309
242
238
217
201
190
180
160
147
123
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 M

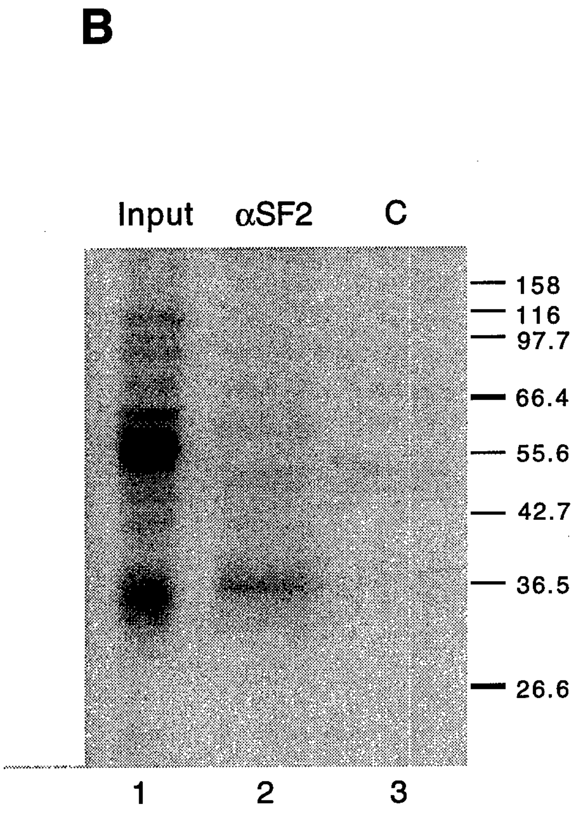
C

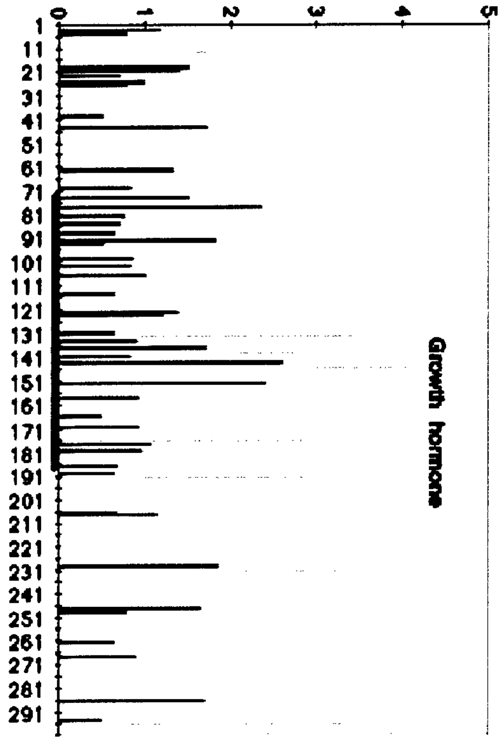

C1
C2
C3
C4
C5
C6
C7
C8
NE
S100
SRp55
+ + + +
+ + + + + + + + + + + +
+ ++ + ++ + ++ + ++
+ + + +
+ + + + + + + + + + + +
+ ++ + ++ + ++ + ++
404
309
242
238
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

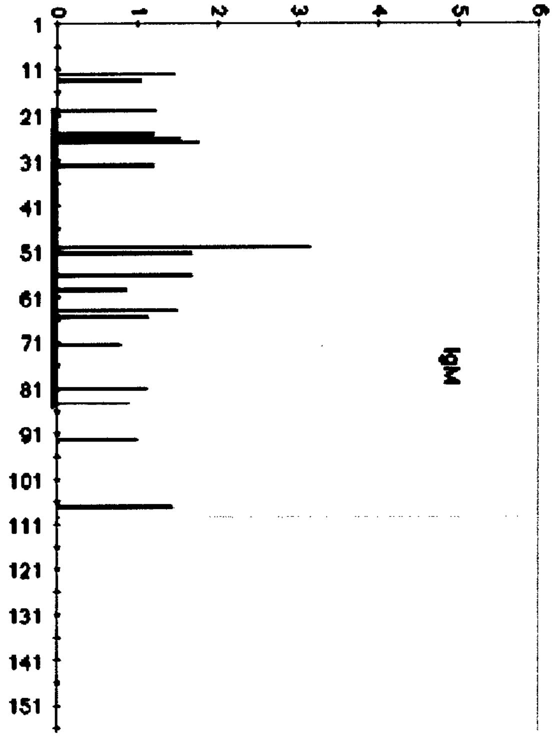

A
E1
E2
E3
E4
E5
E6
E7
SF2/ASF + + + + + + + + + + + + + +
SRp40 + + + + + + + + + + + + + +
404
309
242
238
M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
B
E2
E3
E4
E5
E6
E7
SRp55 + + + + + + + + + + + +
SRp40 + + + + + + + + + + + +
404
309
242
238
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 M

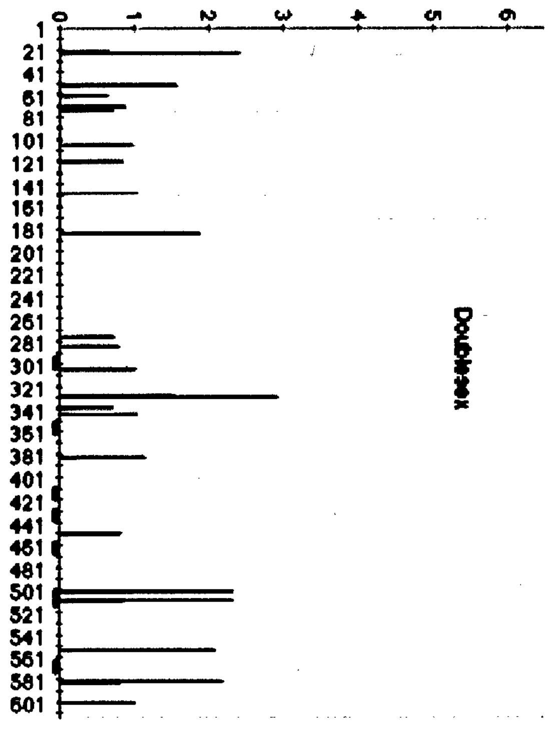
A

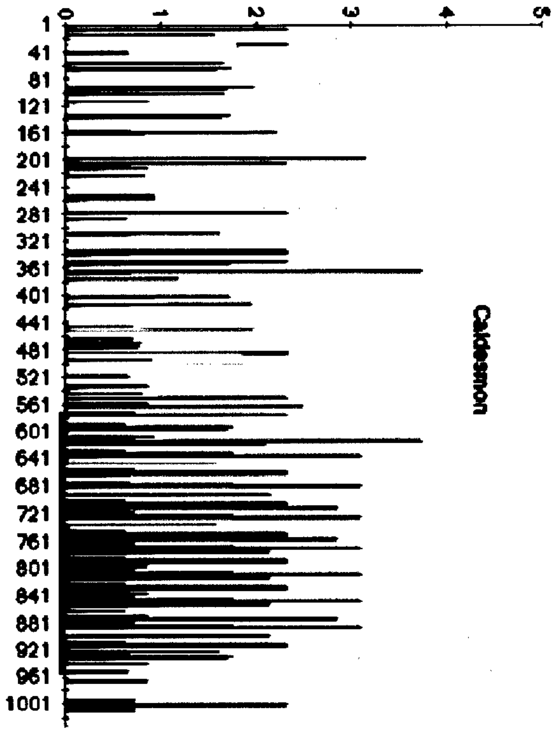

B1E
C4E
A13E
D1E
E7E
G1E
66.4
55.6
42.7
36.5
26.6
1
2
3
4
5
6
7
8
9
10
11
12
13


B
Input
αSF2
C
158
116
97.7
66.4
55.6
42.7
36.5
26.6
1
2
3


Growth hormone
0
1
2
3
4
5
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
251
261
271
281
291


IgM
0
1
2
3
4
5
6
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151


Doublesex
0
1
2
3
4
5
6
1
21
41
61
81
101
121
141
161
181
201
221
241
261
281
301
321
341
361
381
401
421
441
461
481
501
521
541
561
581
601


Caldesmon
0
1
2
3
4
5
1
41
81
121
161
201
241
281
321
361
401
441
481
521
561
601
641
681
721
761
801
841
881
921
961
1001

Acknowledgments

I thank Dr. Y. Shimura for the gift of plasmids μM1-2 and μMΔ, Dr. A. Mayeda for recombinant SR proteins, and other members of our lab for sharing reagents and valuable ideas. I am grateful to Drs. A. Mayeda, M. Murray, L. Cartegni, and D. Horowitz for comments on the manuscript. This work was done in Dr. Adrian Krainer's lab in collaboration with Dr. Michael Zhang, at Cold Spring Harbor Laboratory.

Bibliography

1. **Liu, H.-X.**, Zhang, M., and Krainer, A. R. 1997. Identification of novel classes of exon splicing enhancers recognized by individual SR proteins. Cold Spring Harbor Laboratory RNA Processing Meeting.

2. **Liu, H.-X.**, Zhang, M., and Krainer, A. R. 1997. Identification of Functional Exonic Splicing Enhancer Motifs Recognized by Individual SR Proteins. Manuscript submitted.

In direct cost ($3092):	Cold Spring Harbor Laboratory
"Era of Hope" Meeting ($1050):	Meeting registration, travel, hotel (Renaissance Hotel)
Supplier (~7,000):	The Enzyme Center INC. Pharmacia Biotech INC. Sigma Chemical, CO. Fisher Scientific INC. Dell Marketing LP. Cole Parmer INST. CO. VWR Scientific INC. Boehringer Mannheim Rainin Instruments CO. Mac Warehouse CSHL Bookstore Life Technologies